**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**ĐẠI HỌC KINH TẾ THÀNH PHỐ HỒ CHÍ MINH**

**---**

**NGUYỄN PHẠM DIỆU HIỀN**

**MÔ HÌNH ĐIỂM LỢI NHUẬN ỨNG DỤNG HỌC MÁY TRONG XẾP HẠN THẺ TÍN DỤNG**

**LUẬN VĂN THẠC SĨ KINH TẾ**

TP Hồ Chí Minh – Năm 2024

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**ĐẠI HỌC KINH TẾ THÀNH PHỐ HỒ CHÍ MINH**

**---**

**NGUYỄN PHẠM DIỆU HIỀN**

**MÔ HÌNH ĐIỂM LỢI NHUẬN ỨNG DỤNG HỌC MÁY TRONG XẾP HẠN THẺ TÍN DỤNG**

Chuyên ngành : Tài chính-Ngân hàng (Ngân hàng)

Hướng đào tạo : Hướng ứng dụng

Mã số : 8340201

**LUẬN VĂN THẠC SĨ KINH TẾ**

NGƯỜI HƯỚNG DẪN KHOA HỌC:

TS. VŨ VIỆT QUẢNG

TP Hồ Chí Minh – Năm 2024

Lời cảm ơn

Mục lục

[Lời cảm ơn 3](#_Toc168572143)

[Mục lục 4](#_Toc168572144)

[Danh mục các từ viết tắt 5](#_Toc168572145)

[Danh mục các bảng biểu 6](#_Toc168572146)

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc168572147)

[1.1. Lý do chọn đề tài 1](#_Toc168572148)

[1.2. Mục tiêu nghiên cứu đề bài 1](#_Toc168572149)

[1.3. Đối tượng và phạm vi nghiên cứu 2](#_Toc168572150)

[1.4. Phương pháp nghiên cứu 2](#_Toc168572151)

[1.5. Ý nghĩa nghiên cứu 2](#_Toc168572152)

[1.6. Kết cấu luận văn 3](#_Toc168572153)

[CHƯƠNG 2: TỔNG QUAN VỀ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN 4](#_Toc168572154)

[2.1. Tổng quan về lý thuyết 4](#_Toc168572155)

[2.2. Tổng quan các nghiên cứu trước đây 21](#_Toc168572156)

[CHƯƠNG 3: MÔ HÌNH TÍNH ĐIỂM LỢI NHUẬN 25](#_Toc168572157)

[3.1. Mô tả dữ liệu đầu vào 25](#_Toc168572159)

[3.2. Xây dựng mô hình 38](#_Toc168572165)

[3.3. Chỉ số đánh giá khả năng dự đoán 42](#_Toc168572166)

[CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU 46](#_Toc168572167)

[4.1. Kết quả mô hình 46](#_Toc168572168)

[4.2. Các điểm cần cải thiện 47](#_Toc168572169)

[KẾT LUẬN 48](#_Toc168572170)

[Danh mục tài liệu tham khảo 49](#_Toc168572171)

Danh mục các từ viết tắt

NFI Net Fee Income, doanh thu trừ chi phí liên quan đến phí

NII Net Interest Income, doanh thu trừ chi phí liên quan đến lãi

TOI Total Operating Income, trong bài viết này là tổng NII, NFI

PRO Provision Chi phí dự phòng

TOI-PRO Lợi nhuận sau dự phòng và trước thuế

PBT Profit before tax, Lợi nhuận trước thuế

ADB Dư nợ bình quân trong kỳ của khách hàng

EOP Dư nợ cuối kỳ của khách hàng

PCL Provision for Credit Loss, Chi phí dự phòng

Non-Spender Khách hàng đã được cấp thẻ nhưng chưa sử dụng

Spender Khách hàng được cấp thẻ đã có chi tiêu

Revolver Khách hàng được cấp thẻ đã chi tiêu và có dư nợ phát sinh lãi

Interchange Doanh thu dựa trên các giao dịch từ thẻ

MCBS Chi phí dựa trên các giao dịch từ thẻ

XGBoost Extreme Gradient Boosting, một mô hình học máy

Danh mục các bảng biểu

[Hình 1: Nhóm trường thông tin khách hàng](#Hinh1)

[Hình 2: Đồ thị hàm sigmoid](#Hinh2)

[Hình 3: Thu nhập của ngân hàng theo vòng đời khách hàng](#Hinh3)

[Hình 4: Cấu phần tài chính bình quân năm của thẻ trên Dư nợ trung bình](#Hinh4)

[Bảng 1: Ma trận nhầm lẫn](#Bang3)

[Bảng 2:](#Bang2) Mô tả chi tiết từng nhóm dữ liệu

CHƯƠNG 1: GIỚI THIỆU

* 1. Lý do chọn đề tài

Việc ứng dụng các thuật toán và sự tiện dụng của Học máy trong ngành ngân hàng không còn xa lạ, đặc biệt là với các hoạt động liên quan đến phê duyệt tín dụng thẻ. Ngân hàng bán lẽ trong nước thường có gần 10 nhóm sản phẩm lớn, nhưng thẻ có thể đóng góp lên đến 1/3 thu nhập cho phân khúc này.

Tuy nhiên khác với các sản phẩm tín dụng cho vay (khách hàng ký hợp đồng rõ ràng về khối lượng tiền vay, thời gian trả, lãi suất sẽ trả), ngân hàng không thể ước tính chi phí hoặc lợi nhuận sản phẩm thẻ tín dụng ngay từ khi họ bắt đầu phê duyệt cấp sản phẩm cho khách hàng. Do đó, có 3 lý do ngân hàng có thể hoàn toàn lỗ khi phê duyệt mở thẻ cho khách hàng A mặc dù cơ chế phê duyệt thẻ tín dụng của ngân hàng có thể dự đoán chính xác năng lực tín dụng của khách hàng A:

* Sự phức tạp trong doanh thu và chi phí của thẻ tín dụng (Sẽ được trình bày chi tiết)
* Với thẻ tín dụng, khách hàng không hứa trước thời điểm họ sẽ sử dụng tín dụng và độ lớn khối tín dụng (mặc dù nó sẽ nằm trong ngưỡng tín dụng họ được cấp)
* Việc xét duyệt tín dụng hiện tại chỉ dựa trên cơ sở rủi ro mà khách hàng có khả năng mang lại, chứ không trên cơ sở lợi nhuận mà khách hàng có khả năng mang lại.

Đứng từ góc độ phân tích về tài chính, bài viết sau tập trung vào việc xây dựng và phân tích một mô hình phê duyệt thẻ tín dụng khách hàng trên cơ sở khả năng tạo ra **lợi nhuận** của khách hàng (gọi tắt là mô hình tính điểm lợi nhuận)thay vì đơn thuần là khả năng vỡ nợ của khách hàng như cách các ngân hàng truyền thống đang làm hiện nay.

* 1. Mục tiêu nghiên cứu đề bài

Mục tiêu là xây dựng mô hình lợi nhuận dựa trên năng lực xử lý dữ liệu lớn của Học máy và so sánh với mô hình rủi ro truyền thống, đồng thời đánh giá ảnh hưởng của việc sử dụng mô hình lợi nhuận đối với cấu trúc rủi ro của danh mục thẻ tín dụng.

* 1. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu là hoạt động phê duyệt tín dụng của các tổ chức tài chính quy mô lớn tại Việt Nam, sử dụng dữ liệu từ một ngân hàng thương mại lớn tại Việt Nam để mô phỏng và phân tích

* 1. Phương pháp nghiên cứu

Phần phương pháp nghiên cứu trong luận văn của tôi bắt đầu bằng việc thu thập dữ liệu thực tế từ một nguồn thông tin đáng tin cậy: một tập dữ liệu được chấp nhận từ trước tại một ngân hàng thương mại. Điều này giúp đảm bảo tính chính xác và ứng dụng thực tế của nghiên cứu.

Tiếp theo, tôi tiến hành xây dựng một mô hình máy học đặc biệt để chấm điểm phê duyệt tín dụng cho khách hàng đối với thẻ tín dụng. Trong mô hình này, tôi sử dụng lợi nhuận làm tiêu chí đánh giá thay vì rủi ro, giúp tăng cường hiệu suất của quy trình phê duyệt tín dụng.

Một phần quan trọng của phương pháp là việc so sánh hai danh mục của hai mô hình khác nhau: mô hình hiện tại, dựa trên lợi nhuận, và mô hình truyền thống, dựa trên rủi ro. Qua việc so sánh này, tôi sẽ đánh giá sự hiệu quả và tính ứng dụng của mô hình mới so với phương pháp truyền thống trong việc chấm điểm và phê duyệt tín dụng.

Tóm lại, bài viết sử dụng phương pháp mô phỏng để tạo ra các điểm dữ liệu rủi ro và lợi nhuận, sau đó tiến hành phân tích so sánh giữa hai phương pháp cấp tín dụng dựa trên rủi ro và lợi nhuận.

* 1. Ý nghĩa nghiên cứu

Với việc nghiên cứu sâu hơn về mô hình chấm điểm khách hàng với Học máy dựa trên lợi nhuận, tác giả mong muốn mang lại giá trị:

* Nâng cao hiệu suất phê duyệt tín dụng: Bằng cách sử dụng học máy giúp Ngân hàng có thể đẩy nhanh quá trình phê duyệt tín dụng, giảm thiểu nhân công trong các tác vụ đơn giản
* Giảm tỷ lệ gian lận trong phê duyệt tín dụng: Với Học máy, khả năng gian lận do yếu tố con trong quy trình thẩm định cũng giảm xuống.
* Tối ưu hóa rủi ro và lợi nhuận: Sử dụng lợi nhuận làm tiêu chí đánh giá có thể giúp cân bằng giữa việc giảm thiểu rủi ro tín dụng và tối đa hóa lợi nhuận cho ngân hàng.
* Cải thiện trải nghiệm của khách hàng: Bằng việc sử dụng Học máy trong khâu phê duyệt, Ngân hàng có thể càng gần hơn với giấc mơ tự động hóa toàn bộ các quy trình giúp khách hàng có trải nghiệm tốt hơn, giảm thiểu thời gian chờ đợi và tăng cường sự linh hoạt trong việc cung cấp dịch vụ tài chính.
* Đóng góp cho sự phát triển của ngành tài chính: Nghiên cứu này có thể đưa ra các phương pháp và tiêu chuẩn mới cho việc đánh giá tín dụng trong ngành tài chính, góp phần vào sự tiến bộ và sự phát triển của lĩnh vực này.
  1. Kết cấu luận văn

Do mô hình dựa vào nền tảng Học máy, bài viết được tách thành 3 phần, giải thích nhiều về các kỹ thuật xử lý cơ bản của Học máy và các đặc trưng về thẻ tín dụng giúp người đọc từ góc độ tài chính có thể dễ nắm bắt. Cụ thể:

Chương 1: Cơ sở lý thuyết về xếp hạn tín dụng và Lý thuyết về mô hình Học máy

Chương 2: Xây dựng mô hình tính điểm lợi nhuận

Giải thích các câu hỏi:

* Mô hình được xây dựng và hoạt động như thế nào
* Độ tin cậy của mô hình

Chương 3: Sự ưu việt và hạn chế của mô hình so với mô hình tính điểm truyền thống

* Với tập khách hàng cutoff tại điểm tốt nhất như trên profit score thì risk probility thì hệ số rủi ro sẽ như thế nào

Chương 4: Giải pháp hoàn thiện mô hình có thể đạt được trong tương lai

CHƯƠNG 2: TỔNG QUAN VỀ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

* 1. Tổng quan về lý thuyết
     1. Tổng quan về xếp hạng tín dụng khách hàng cá nhân
        1. Khái niệm

Xếp hạng tín dụng khách hàng cá nhân là một hệ thống đánh giá được các tổ chức tài chính, như ngân hàng hoặc tổ chức tín dụng, sử dụng để đánh giá rủi ro tín dụng của mỗi khách hàng cá nhân. Thông thường, mục tiêu của việc xếp hạng này là đưa ra dự đoán về khả năng của một cá nhân trong việc hoàn trả nợ và thực hiện các cam kết tài chính trong tương lai.

* + - 1. Vai trò của xếp hạng tín dụng khách hàng cá nhân đổi với Ngân hàng

Quản lý rủi ro: Xếp hạng tín dụng giúp ngân hàng đánh giá và quản lý rủi ro tín dụng của các khoản vay. Bằng cách đánh giá rủi ro của từng khách hàng, ngân hàng có thể xác định liệu họ có thể trả nợ đúng hạn hay không.

Quyết định về cấp tín dụng: Dựa trên xếp hạng tín dụng, ngân hàng quyết định cấp mức tín dụng (như thẻ tín dụng, khoản vay cá nhân) và điều kiện (lãi suất, hạn mức) cho từng khách hàng. Các khách hàng có xếp hạng tốt có thể được cấp tín dụng với lãi suất thấp hơn và hạn mức cao hơn.

Chính sách giải quyết nợ: Xếp hạng tín dụng cũng giúp ngân hàng quyết định cách tiếp cận và xử lý các khoản nợ không trả được của khách hàng. Các khách hàng có rủi ro tín dụng cao hơn thường phải chịu các biện pháp thu hồi nợ nghiêm ngặt hơn.

Bảo vệ lợi ích của ngân hàng: Bằng cách đánh giá rủi ro tín dụng của từng khách hàng, ngân hàng có thể đảm bảo rằng họ đang cung cấp các sản phẩm và dịch vụ tài chính một cách có trách nhiệm, đồng thời bảo vệ lợi ích của chính họ khỏi các khoản nợ không trả được và mất mát tài chính.

* + - 1. Sự cần thiết của xếp hạng tín dụng khách hàng cá nhân

Trong trưng hợp một ngân hàng không có hệ thống đánh giá điểm tín dụng hoặc quy trình đánh giá điểm tín dụng có vấn đề thì:

Tăng nguy cơ rủi ro tín dụng: Không có hệ thống đánh giá tín dụng hoặc quy trình đánh giá không hiệu quả có thể dẫn đến việc cung cấp tín dụng cho những khách hàng có khả năng không trả nợ hoặc có rủi ro tín dụng cao hơn mà không được xác định rõ. Điều này có thể dẫn đến tăng nguy cơ cho ngân hàng về các khoản nợ không trả được và mất mát tài chính.

Mất lợi thế cạnh tranh: Một hệ thống đánh giá tín dụng hiệu quả giúp ngân hàng thu hút và duy trì khách hàng có rủi ro tín dụng thấp, từ đó giảm nguy cơ rủi ro và cung cấp sản phẩm và dịch vụ với điều kiện tốt hơn. Nếu không có hoặc có vấn đề với quy trình đánh giá, ngân hàng có thể mất lợi thế cạnh tranh trước các đối thủ có hệ thống đánh giá tín dụng tốt hơn.

Tăng chi phí hoạt động: Việc không có hệ thống đánh giá tín dụng hoặc có vấn đề với quy trình đánh giá có thể dẫn đến việc ngân hàng phải xử lý nhiều khoản nợ không trả được hơn, đồng thời tăng chi phí thu hồi nợ và rủi ro về mất mát tài chính.

Tác động đến uy tín: Nếu ngân hàng không thể đánh giá và quản lý rủi ro tín dụng một cách hiệu quả, điều này có thể ảnh hưởng tiêu cực đến uy tín và hình ảnh của họ trong mắt cả khách hàng và các cơ quan giám sát tài chính.

* + - 1. Quy định liên quan

Quy định về xếp hạng tín dụng khách hàng cá nhân của Ngân hàng nhà nước:

“Theo Quy định của Ngân hàng Nhà nước Việt Nam: tối đa đến 07/05/2008 các Ngân hàng phải xây dựng hệ thống xếp hạng tín dụng nội bộ để hỗ trợ cho việc phân loại nợ, quản lý chất lượng tín dụng phù hợp với phạm vi hoạt động, tình hình thực tế của Ngân hàng và phải đáp ứng đủ các điều kiện như sau:

* Hệ thống xếp hạng tín dụng đã được áp dụng thử nghiệm tối thiểu 1 năm.
* Kết quả xếp hạng tín dụng được Hội đồng quản trị phê duyệt.
* Hệ thống xếp hạng tín dụng nội bộ phù hợp với hoạt động kinh doanh, đối tượng khách hàng, tính chất rủi ro của khoản nợ của tổ chức tín dụng.
* Chính sách quản lý rủi ro tín dụng, mô hình giám sát rủi ro tín dụng, phương pháp xác định và đo lường rủi ro tín dụng có hiệu quả, trong đó, bao gồm cách thức đánh giá về khả năng trả nợ của khách hàng, hợp đồng tín dụng, các loại tài sản bảo đảm, khả năng thu hồi nợ và quản lý nợ của tổ chức tín dụng.
* Phân định rõ ràng trách nhiệm và quyền hạn của Hội đồng quản trị, Tổng giám đốc trong việc phê duyệt, thực hiện và kiểm tra thực hiện hệ thống xếp hạng tín dụng và chính sách dự phòng của tổ chức tín dụng và tính độc lập của các bộ phận quản lý rủi ro.
* Hệ thống thông tin có hiệu quả để đưa ra các quyết định, điều hành và quản lý đối với hoạt động kinh doanh của tổ chức tín dụng và thích hợp với hệ thống xếp hạng tín dụng và phân loại nợ.

***Hệ thống xếp hạng tín dụng nội bộ tối thiểu phải bao gồm***:

* Các cơ sở pháp lý liên quan đến thành lập và ngành nghề kinh doanh của khách hàng.
* Các chỉ tiêu kinh tế tổng hợp liên quan đến tình hình kinh doanh, tài chính, tài sản, khả năng thực hiện nghĩa vụ theo cam kết.
* Uy tín với các tổ chức tín dụng đã giao dịch trước đây.
* Các chỉ tiêu đánh giá khách hàng chi tiết, cụ thể, có hệ thống (đánh giá yếu tố ngành nghề và địa phương) trên cơ sở đó xếp hạng có thể đối với khách hàng.

Dựa trên các yêu cầu trên, Ngân hàng xây dựng hệ thống xếp hạng tín dụng nội bộ, trình Ngân hàng Nhà nước chính sách dự phòng rủi ro và chỉ được thực hiện sau khi Ngân hàng Nhà nước chấp thuận bằng văn bản.”

* + 1. Tổng quan về mô hình học máy
       1. Khái niệm và lịch sử phát triển Học máy

Theo Murphy, trích ấn phẩm Đại học MIT năm 2012 tiêu đề "Machine Learning: A Probabilistic Perspective":

“We are entering the era of big data. For example, there are about 1 trillion web pages; one hour of video is uploaded to YouTube every second, amounting to 10 years of content every day; the genomes of 1000s of people, each of which has a length of 3.8 × 109 base pairs, have been sequenced by various labs; Walmart handles more than 1M transactions per hour and has databases containing more than 2.5 petabytes (2.5 × 1015) of information (Cukier 2010); and so on.

This deluge of data calls for automated methods of data analysis, which is what machine learning provides. In particular, we define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!)”

Murphy đưa ra một định nghĩa bao quát về Machine Learning, xác định nó như một tập hợp các phương pháp tự động tìm ra quy luật của dữ liệu và từ đó dự đoán xu hướng tương lai của dữ liệu đó. Điểm tập trung của nó là phát triển các hệ thống có khả năng tự học và cải thiện qua thời gian mà không cần phải lập trình cụ thể. Trong lĩnh vực này, các mô hình và thuật toán được sử dụng để phân tích dữ liệu, nhận biết mẫu và đưa ra dự đoán hoặc quyết định mà không cần phải tuân theo hướng dẫn lập trình trước.

Trong ngành ngân hàng, việc tiếp nhận và phân tích dữ liệu khách hàng để dự đoán hành vi trong quy trình phê duyệt thẻ tín dụng là một thách thức lớn. Đối mặt với lượng lớn thông tin từ nhiều nguồn khác nhau, từ thông tin cá nhân của khách hàng đến dữ liệu từ các tổ chức tín dụng khác và từ Big Data, việc này trở nên phức tạp hơn bao giờ hết. Với việc số lượng khách hàng ngày càng tăng, việc xử lý mỗi hợp đồng trở nên quá tải và đòi hỏi sự hỗ trợ mạnh mẽ từ các công cụ Machine Learning.

Lịch sử của việc ứng dụng Học máy trong nhàng ngân hàng đã nhen nhóm xuất hiện từ những năm 1980. Năm 1982, Apex đã tạo ra PlanPower, một chương trình trí tuệ nhân tạo dành cho tư vấn thuế và tài chính được cung cấp cho khách hàng thượng lưu. Năm 1987, Ngân hàng Chase Lincoln First (nay là một phần của JP Morgan Chase) đã ra mắt Hệ thống Lập kế hoạch Tài chính Cá nhân. Ngay sau đó, vào năm 1989, Điểm tín dụng FICO, một công thức xếp hạng tín dụng khách hàng cá nhân dựa trên một thuật toán tương tự được các ngân hàng sử dụng ngày nay, đã được phát hành.

Ngày này, việc ứng dụng Học máy trong quá trình phê duyệt tín dụng cá nhân lan khắp hầu hết các tổ chức tín dụng trên thế giới. Tại Việt Nam, ngoài các tổ chức tín dụng đa quốc gia như (Home Credit, Standard Charter, HSBC, …), một số ngân hàng nội địa dẫn đầu thị trường đã bắt đầu ứng dụng Học máy trong một phần trong quy trình phê duyệt tín dụng, thường là trong lúc xây dựng mô hình tính điểm tín dụng. Ngoài ra, các thuật toán để giải bài toán về sản phẩm tín dụng cho khách hàng cá nhân cũng được các ngân hàng áp dụng như:

* VPBank sử dụng thuật toán Học máy mãng Natural Language Processing để nhận diện giao dịch tìm ẩm rủi ro và phân loại phân khúc khách hàng
* TPBank đã tích hợp công nghệ nhận diện khuôn mặt vào kênh ngân hàng tự động LiveBank của mình, tăng cường tính bảo mật và tiện lợi cho khách hàng.
* VietinBank sử dụng các Học máy với công nghệ nhận diện FaceID để xác định khách hàng và chuyển tiếp yêu cầu của họ đến các tư vấn viên.
* Các ngân hàng khác như VietABank, Nam A Bank, Techcombank, VIB và ACB đã áp dụng trí tuệ nhân tạo trong các chức năng khác nhau, bao gồm chatbot hỗ trợ và tương tác khách hàng, quản lý tài sản, bảo mật, phòng chống gian lận và phân tích việc rút tiền ATM trong mùa cao điểm.
  + - 1. Đặc trưng heuristic và black-box của học máy

Đặc trưng heuristic của một mô hình học máy đề cập đến khả năng của mô hình đó trong việc tự học và tự cải thiện thông qua kinh nghiệm, thay vì dựa vào các nguyên tắc hoặc quy tắc cụ thể. Điều này có nghĩa là mô hình học máy có khả năng tìm ra các giải pháp tối ưu thông qua việc thử nghiệm và đánh giá các kết quả, mà không cần phải tuân theo các quy tắc cứng nhắc. Ví dụ: Thông thường để dự đoán giá cổ phiếu, nhà đầu tư sử dụng các công thức tính toán dựa trên các chỉ số tài chính và kinh doanh để dự đoán giá cổ phiếu trong tương lai dựa trên một hàm tuyến tính để kết hợp các yếu tố như lợi nhuận tăng trưởng hàng năm và tỷ lệ lợi nhuận gộp để dự đoán giá cổ phiếu. Trong khi đó, một mô hình Học máy nặng tính heuristic như một mạng nơ-ron hồi quy (RNN) hoặc mạng nơ-ron tái lập (LSTM) có thể học từ dữ liệu lịch sử về giá cổ phiếu và các yếu tố liên quan. Mô hình này có thể phát hiện ra các mẫu phức tạp và tương quan không rõ ràng giữa các yếu tố và giá cổ phiếu. Ví dụ, nó có thể nhận ra rằng khi lợi nhuận tăng cao trong quý trước, giá cổ phiếu có xu hướng tăng trong quý tiếp theo, hoặc rằng sự thay đổi trong giá dầu có thể ảnh hưởng đến giá cổ phiếu của các công ty trong ngành năng lượng

Dưới đây là một số điểm quan trọng liên quan đến tính heuristic của mô hình học máy:

* Tự học: Mô hình học máy có khả năng tự điều chỉnh và cải thiện qua thời gian khi được huấn luyện trên dữ liệu mới. Thay vì phải được lập trình để thực hiện một nhiệm vụ cụ thể, chúng có khả năng tự động học từ dữ liệu và điều chỉnh các tham số để tối ưu hóa hiệu suất.
* Khả năng tự điều chỉnh: Mô hình học máy có thể tự điều chỉnh các tham số hoặc cấu trúc của chính nó dựa trên dữ liệu mẫu. Điều này cho phép chúng thích nghi với các biến đổi trong dữ liệu hoặc môi trường mà không cần sự can thiệp từ con người.
* Tính linh hoạt: Mô hình học máy có thể áp dụng cho nhiều loại vấn đề khác nhau và có khả năng tìm ra các giải pháp tốt nhất cho mỗi vấn đề cụ thể. Điều này cho phép chúng được sử dụng trong nhiều lĩnh vực khác nhau, từ dự báo tài chính đến nhận dạng hình ảnh.
* Tính chủ động: Mô hình học máy có khả năng tự động đưa ra các quyết định dựa trên dữ liệu mà không cần sự can thiệp hoặc hướng dẫn từ con người. Điều này giúp chúng thích nghi nhanh chóng với các tình huống mới và không rõ ràng.

Nhiều mô hình học máy thường dựa trên cách tiếp cận heuristics và không đặt ra giả định rõ ràng về các mối quan hệ giữa các biến của mô hình. Mặc dù điều này khiến chúng linh hoạt và dễ đưa ra dự đoán tốt hơn, nhưng lại khiến các mô hình này khó giải thích về kết quả mà chính mô hình đưa ra, gây ra vấn đề gọi là vấn đề black-box. Ngoài ra, điều này cũng là một trong những nguyên nhân dẫn đến vấn đề phổ biến trong các bài toán ứng dụng Học máy là quá khớp (overfitting[[1]](#footnote-1)).

* + - 1. Kỹ thuật chấm điểm quan trọng của biến đầu vào

Việc xây đựng một mô hình có rất nhiều các biến đầu vào không có ý nghĩa với mô hình, tức dù biến đầu vào có hay không thì mô hình vẫn trả ra những giá trị không thay đổi mang lại một số rủi ro nhưng khó khăn trong quyết định liệu có cần lượt bỏ điểm dữ liệu khi bị thiếu dữ liệu tại trường đó thì các biến đầu vào không có ý nghĩa còn làm mô hình tốn nguồn lực máy chạy hơn một cách rõ ràng khi dữ liệu đầu vào rất lớn. Do đó, có một số kỹ thuật giúp chấm điểm quan trọng của biến đầu vào (Feature Importance Score) như:

**Mean Decrease in Impurity (hoặc Gini Importance) – Điểm quan trọng mặc định**

Phương pháp mặc định để tính đo lường sự quan trọng của các biến trong mô hình là dựa trên cơ chế "mean decrease in impurity" hoặc còn được gọi là "Gini importance". Ở mỗi bước chia trong mỗi cây, sự cải thiện trong chỉ số chia (split-criterion) được tính là đo lường quan trọng của biến chia, và được tích lũy qua tất cả các cây trong rừng riêng biệt cho từng biến. Một cách khác để hiểu, ở mỗi nút chia trong mỗi cây, thuật toán tính toán mức độ làm giảm sự không thuần túy (impurity) hoặc lỗi Gini khi chia dữ liệu bằng mỗi biến. Các biến mà khi chia dẫn đến sự giảm lớn nhất trong không thuần túy hoặc lỗi Gini được coi là quan trọng hơn. Phương pháp này tương tự như cách 𝑅2 được tính trong hồi quy tuyến tính khi đo lường phần trăm biến thiên của biến phụ thuộc mà mô hình có thể giải thích được. Tương tự, đo lường sự quan trọng của biến trong mô hình cây quyết định là mức độ mà biến có thể giải thích sự biến thiên trong dữ liệu, được tính bằng cách đánh giá cải thiện của chỉ số chia khi sử dụng biến đó.

**Permutation Feature Importance – Điểm quan trọng hoán vị**

Điểm quan trọng hoán vị được Breiman lần đầu giới thiệu vào năm 2001 về thuật toán Random Forest. Sau đó được nâng cấp bởi Fisher, Rudin, và Dominici vào năm 2018 với một phiên bản không phụ thuộc vào mô hình cụ thể với cái tên Mô hình phụ thuộc (Model Reliance[[2]](#footnote-2)). Mô hình này đo độ quan trọng của một biến đầu vào bằng cách đo độ tăng lỗi trong dự đoán sau mỗi khi hoán đổi biến đầu vào. Tóm lại, biến đầu vào được cho là quan trọng khi việc hoán đổi biến làm tăng số lượng lỗi trong dự đoán của mô hình.

* + - 1. Mô hình Học máy - Logistic regression:

**Nguyên tắc cơ bản**

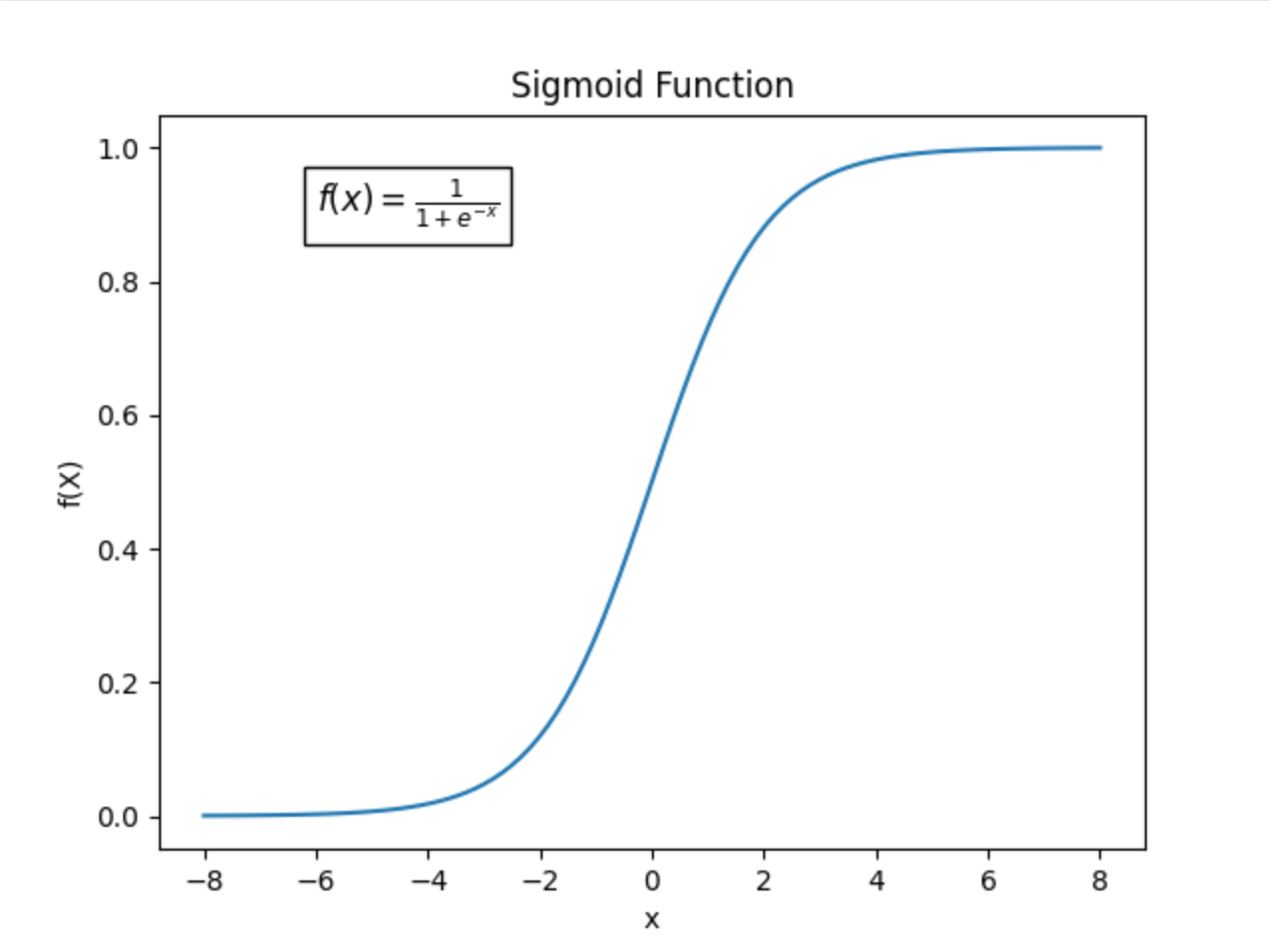
Trong bài toán dự đoán một giá trị với hai lựa chọn (Ví dụ như dự đoán khách hàng có vỡ nợ hoặc không vỡ nợ, liệu giao dịch này là gian lận hay không gian lận,…), người ta thường dùng đến bài Hồi quy Logistic.

Một cách tổng quan, với Hồi quy logicstic, ngân hàng mặc dù chỉ có khả năng truyền vào thông tin liệu khách hàng A có khả năng vỡ nợ hay không, tức 0 - Negative hoặc 1-Positive trong tập dữ liệu huấn luyện mô hình. Nhưng kết quả trả ra của mô hình cho biết liệu xác xuất vỡ nợ của khách hàng là bao nhiêu.

**Hàm Hồi quy Logistic:**

**Hình 2: Đồ thị hàm sigmoid**

Với x =



Một cách tiếp cận khác, Odds (tỷ lệ cơ hội), đề cập đến tỷ lệ giữa xác suất của sự kiện xảy ra và xác suất của sự kiện không xảy ra. Nói cách khác, odds là tỷ lệ giữa khả năng thành công và khả năng thất bại trong một tình huống. Với Odds là tỷ lệ xác suất số mẫu trường hợp không vỡ nợ (positive với y = 1) trên tỷ lệ xác suất số mẫu trường hợp vỡ nợ (negative với y = 0)

Do đó hàm Hồi quy Logistic còn có thể viết lại như sau:

Với

**Kiểm định mô hình:**

Trong kiểm định mô hình Hồi quy Logistic, người ta thường sử dụng Ma trận nhầm lẫn (Confusion matrix).

**Bảng 1: Ma trận nhầm lẫn**

|  |  |  |
| --- | --- | --- |
|  | Dự báo = 0 | Dự báo = 1 |
| Thực tế = 0 | TN – True Negative | FP - False Positive |
| Thực tế = 1 | FN - False Negative | TP – True Positive |

**Độ nhạy (Sensitive)** còn được gọi là "tỉ lệ True Positive" (TPR), là một thước đo cho biết khả năng của mô hình trong việc phát hiện các trường hợp thực tế không vỡ nợ (1).

**Công thức tính độ nhạy là:**

**Sensitive**

**Trong đó:**

* TP (True Positive) là số lượng các trường hợp mà mô hình dự đoán đúng là positive (1).
* FN (False Negative) là số lượng các trường hợp mà mô hình dự đoán là negative (0) nhưng thực tế là positive (1).

Độ nhạy càng cao thì mô hình càng tốt trong việc phát hiện các trường hợp positive. Nói cách khác, nếu độ nhạy gần 1, có nghĩa là mô hình có khả năng phát hiện được hầu hết các trường hợp positive.

Tóm lại, trong các bài toán về rủi ro vỡ nợ khách hàng thường sẽ được sử dụng mô hình Hồi quy Logistic. Một phần vì năng lc giải thích của mô hình đối với kết quả mà nó đã đưa ra, một phần khác vì kết quả của mô hình luôn nằm trong giới hạn [0-1] giúp nhà phân tích dễ dàng xác định khả năng vỡ nợ của khách hàng.

* + - 1. Mô hình Học máy - Random Forest

Random Forest là một thuật toán học máy thuộc về nhóm supervised learning, có thể được sử dụng cho cả bài toán phân loại (classification) và hồi quy (regression). Thuật toán này được phát triển dựa trên việc kết hợp nhiều cây quyết định (decision trees) để cải thiện hiệu suất và giảm nguy cơ overfitting.

Nguyên Lý Hoạt Động

* *Tạo Rừng (Forest Creation)*

1. Bootstrap Aggregating (Bagging): Để tạo ra một rừng cây (forest), Random Forest sử dụng kỹ thuật bootstrap, tức là chọn ngẫu nhiên các mẫu từ tập dữ liệu gốc với phép chọn lại (sampling with replacement) để tạo thành nhiều tập dữ liệu con (subsets). Mỗi tập con này sẽ được sử dụng để huấn luyện một cây quyết định.
2. Random Feature Selection: Khi xây dựng mỗi cây quyết định, tại mỗi nút (node) của cây, một tập con ngẫu nhiên của các đặc trưng (features) sẽ được chọn thay vì sử dụng toàn bộ các đặc trưng. Điều này giúp tạo ra sự đa dạng giữa các cây trong rừng.

* *Dự Đoán (Prediction)*

1. Phân Loại (Classification): Đối với bài toán phân loại, mỗi cây quyết định trong rừng sẽ đưa ra một dự đoán về nhãn của mẫu. Kết quả cuối cùng là nhãn được dự đoán nhiều nhất (majority vote) từ các cây.
2. Hồi Quy (Regression): Đối với bài toán hồi quy, mỗi cây quyết định đưa ra một giá trị dự đoán, và kết quả cuối cùng là giá trị trung bình của tất cả các dự đoán từ các cây.

* *Ưu Điểm*

1. Giảm Overfitting: Bằng cách kết hợp nhiều cây quyết định, Random Forest giảm nguy cơ overfitting so với việc sử dụng một cây đơn lẻ.
2. Ổn Định và Chính Xác: Random Forest thường có hiệu suất tốt và ổn định trên nhiều tập dữ liệu khác nhau.
3. Xử Lý Tốt Dữ Liệu Lớn và Phức Tạp: Thuật toán này có thể xử lý tập dữ liệu lớn và phức tạp với nhiều đặc trưng khác nhau.
4. Đánh Giá Tầm Quan Trọng của Đặc Trưng: Random Forest cung cấp thông tin về tầm quan trọng của từng đặc trưng, giúp xác định các đặc trưng quan trọng nhất đối với mô hình.

* *Nhược Điểm*

1. Tính Tốn Kém về Tính Toán: Việc xây dựng và dự đoán với nhiều cây quyết định có thể đòi hỏi tài nguyên tính toán lớn.
2. Giảm Khả Năng Giải Thích: So với một cây quyết định đơn giản, Random Forest khó giải thích hơn do sự phức tạp của việc kết hợp nhiều cây.

Hyperparameter:

* n\_estimators: Đây là số lượng cây quyết định trong Random Forest. Nó xác định số lượng cây sẽ được xây dựng trong quá trình huấn luyện. Giá trị này cần được lựa chọn sao cho đủ lớn để mô hình có khả năng học được các mẫu phức tạp, nhưng cũng không quá lớn để tránh tăng quá mức độ phức tạp và thời gian huấn luyện.
* max\_features: Đây là số lượng đặc trưng (features) được chọn ngẫu nhiên để xem xét khi tạo ra một cây quyết định. Giá trị này xác định số lượng đặc trưng tối đa mà mỗi cây có thể sử dụng để đưa ra quyết định. Việc giới hạn số lượng đặc trưng có thể giúp tránh overfitting và tăng tính đa dạng của các cây trong Random Forest.
* max\_depth: Đây là độ sâu tối đa của mỗi cây quyết định trong Random Forest. Nếu giá trị max\_depth được xác định, các cây sẽ không được phép phân chia thêm khi đạt đến độ sâu này. Giới hạn độ sâu của cây có thể giúp kiểm soát độ phức tạp của mô hình và tránh overfitting.
* min\_samples\_split: Đây là số lượng mẫu tối thiểu yêu cầu để một nút trong cây có thể được phân chia thành các nhánh con. Giá trị này giúp kiểm soát quá trình phân tách và tránh việc tạo ra các nhánh con quá nhỏ, không có độ tương tự đáng kể.
* min\_samples\_leaf: Đây là số lượng mẫu tối thiểu yêu cầu trong mỗi lá (leaf) của cây quyết định. Nếu giá trị này được xác định, các cây sẽ không được phép phân chia thành lá nhỏ hơn với số lượng mẫu ít hơn giá trị min\_samples\_leaf. Giới hạn số lượng mẫu tối thiểu trong lá có thể giúp tránh overfitting và tạo ra các lá có độ tương tự đáng kể.
  + - 1. Mô hình Học máy - XGBoost

XGBoost - Extreme Gradient Boosting, đã được sử dụng rộng rãi trong nhiều lĩnh vực để đạt được các kết quả hàng đầu trong một số thách thức dữ liệu (ví dụ, các cuộc thi Kaggle), đây là một hệ thống học máy có hiệu suất cao, có khả năng mở rộng và hiệu quả cho việc tăng cường cây. XGBoost được tối ưu hóa trong khung Gradient Boosting và được phát triển bởi Chen và Guestrin vào năm 2016. Sau đó, nó liên tục được sử dụng trong các nghiên cứu phát triển mô hình dành cho lĩnh vực ngân hàng như Liu và Zhang vào 2022, Baesens và Smedts vào 2023.

XGBoost thuộc loại thuật toán boosting, nơi các mô hình yếu (thường là cây quyết định) được xây dựng và kết hợp lại thành một mô hình mạnh mẽ hơn. Mỗi cây quyết định trong XGBoost là một chuỗi các quyết định logic được thực hiện theo các biến đầu vào để dự đoán kết quả. Cây quyết định được xây dựng theo cách tối ưu hóa một hàm mất mát (loss function), thông qua việc chia nhỏ dữ liệu thành các nhóm ngày càng nhỏ sao cho việc phân loại được tối ưu.

XGBoost sử dụng phương pháp gradient boosting để cải thiện các cây quyết định. Ý tưởng chính là mỗi cây mới được xây dựng để sửa các lỗi của cây trước đó. Đối với mỗi lần lặp (iteration), XGBoost tính toán gradient của hàm mất mát (loss function) để tìm cây quyết định tiếp theo có thể giảm lỗi dự đoán. Sau đó, cây quyết định mới được thêm vào mô hình và được cân nhắc dựa trên mức độ cải thiện trong dự đoán. Do đặc tính này người ta gọi nó với cái tên là “Gradient Boosting”.

Trong XGBoost, các mô hình con trả ra kết quả và kiểm định chéo với nhau. Quá trình này gọi là Cross validation. Trong đó, cơ sở dữ liệu mô phỏng được chia ngẫu nhiên thành tập huấn luyện (80%) và tập kiểm tra (20%). Phương pháp kiểm tra chéo K-fold (số mẫu là k mẫu) được sử dụng cho tập huấn luyện để tìm ra các tham số tốt nhất cho mô hình. Trong kiểm tra chéo 10-fold, tập huấn luyện được chia thành mười phần có kích thước gần như bằng nhau, trong đó chín phần được sử dụng để huấn luyện và một phần được sử dụng để kiểm tra. Quá trình này được lặp lại mười lần theo chu kỳ và trung bình của các độ chính xác này được sử dụng như là độ chính xác dự đoán dự kiến.

XGBoost cũng sử dụng kỹ thuật regularization để kiểm soát việc tạo ra các cây quyết định quá phức tạp, tránh overfitting. Regularization có thể thực hiện thông qua các siêu tham số như learning rate, max\_depth, min\_child\_weight, gamma, lambda (L2 regularization term), và alpha (L1 regularization term).

Công thức:

Tập huấn luyện đầu vào: , với hàm mất mát (loss function) 𝐿(𝑦,𝐹(𝑥)), Số lượng weak learners và tốc độ học (Learning rate) 𝛼. Mô hình đi qua các bước như sau:

Bước 1. Mô hình gốc với hằng số

Bước 2: Vòng lặp m chạy từ [1:M]

Bước 3: Trả kết quả đầu ra

Kiểm định mô hình:

XGBoost có thể được cấu hình với nhiều hàm mất mát khác nhau tùy thuộc vào loại vấn đề, ví dụ: hồi quy (regression), phân loại nhị phân (binary classification), và phân loại đa lớp (multiclass classification). Một số hàm mất mát phổ biến trong XGBoost bao gồm Mean Squared Error (MSE) cho hồi quy và Logarithmic Loss (logloss) cho phân loại.

Đặc tính

Khác với Hồi quy Logictics, XGBoost quá phức tạp với nhiều tầng và lớp mô hình con dẫn đến việc kết quả của một mô hình lớn XGBoost dẫn đến việc không có nhiều khả năng giải thích với chính kết quả mà nó đưa ra. Tuy nhiên, bù lại XGBoost ít bị vấn đề quá khớp (Overfit) mà các mô hình đơn giản như Hồi quy Logistic có thể gặp phải.

Hyperparameter:

Thông thường, các hệ số là hằng số được điều chỉnh riêng lẻ tùy thuộc vào từng tập huân luyện như sau:

* learning\_rate: Đây là tỷ lệ học (learning rate) được sử dụng để điều chỉnh đóng góp của mỗi cây trong quá trình huấn luyện. Giá trị learning\_rate càng nhỏ, đóng góp của mỗi cây càng nhỏ, dẫn đến tốc độ học chậm hơn nhưng có thể giúp tránh việc mô hình quá khớp. Ngược lại, giá trị learning\_rate càng lớn, đóng góp của mỗi cây càng lớn, tăng tốc độ học nhưng có thể gây hiện tượng quá khớp.
* n\_estimators: Đây là số lượng cây quyết định trong ensemble (tổ hợp) của XGBoost. N\_estimators quyết định độ phức tạp của mô hình và số lượng cây cần huấn luyện. Tăng giá trị n\_estimators có thể cải thiện hiệu suất của mô hình, nhưng đồng thời cũng tăng thời gian huấn luyện.
* max\_depth: Đây là chiều sâu tối đa của mỗi cây quyết định trong XGBoost. Giới hạn chiều sâu có thể giúp kiểm soát độ phức tạp của cây và tránh overfitting. Nếu max\_depth quá lớn, mô hình có thể bị overfitting và khó được tổng quát hoá. Ngược lại, nếu max\_depth quá nhỏ, mô hình có thể bị underfitting và không thể học được các quy tắc phức tạp.
* subsample: Đây là tỷ lệ mẫu con (subsample) được sử dụng trong quá trình huấn luyện từ tập dữ liệu huấn luyện. Giá trị subsample nhỏ hơn 1.0 cho phép chọn ngẫu nhiên một phần của dữ liệu để huấn luyện các cây, giúp mô hình tránh overfitting và tăng tính tổng quát hóa.
* colsample\_bytree và colsample\_bylevel: Đây là tỷ lệ các biến đặc trưng (features) được sử dụng cho mỗi cây (colsample\_bytree) và mỗi mức (level) trong cây (colsample\_bylevel). Các giá trị nhỏ hơn 1.0 cho phép chọn ngẫu nhiên một phần của các biến đặc trưng, giúp tránh overfitting và tăng khả năng tổng quát hóa.
* reg\_alpha và reg\_lambda: Đây là các tham số điều chuẩn (regularization) được sử dụng để kiểm soát độ phức tạp của mô hình và tránh overfitting. Reg\_alpha là hệ số điều chuẩn L1, trong khi reg\_lambda là hệ số điều chuẩn L2. Giá trị reg\_alpha và reg\_lambda càng lớn, độ điều chuẩn càng mạnh.
* Gamma: Trong mô hình XGBoost, gamma là một siêu tham số (hyperparameter) quan trọng để điều chỉnh quá trình cắt tỉa (pruning) của cây quyết định trong quá trình huấn luyện. Gamma xác định mức tăng tối thiểu cần thiết để tiếp tục chia nhánh một nút trong cây. Khi xây dựng cây quyết định, thuật toán XGBoost sẽ đánh giá mức độ giảm tối đa (maximum reduction) của hàm mất mát (loss function) khi tiến hành chia nhánh. Nếu mức giảm này nhỏ hơn giá trị gamma, quá trình chia nhánh sẽ dừng lại, và nút hiện tại sẽ được coi là một nút lá (leaf node) không tiếp tục phân chia. Giá trị gamma càng lớn, quá trình cắt tỉa sẽ càng được áp dụng mạnh mẽ hơn, giúp giảm độ phức tạp của cây và chống lại hiện tượng quá khớp (overfitting). Tuy nhiên, giá trị gamma quá cao có thể dẫn đến mất mát thông tin quan trọng và gây hiệu suất kém trên tập dữ liệu kiểm tra.
  + - 1. Weight of Evidence (WOE)

WOE là một phương pháp để đo lường sức mạnh phân biệt của một biến số với kết quả mong muốn (như việc khách hàng trả nợ đúng hạn hay không). Nó đưa ra độ đo về mức độ mà một giá trị của biến số có khả năng liên quan đến kết quả mong muốn. Công thức tính WOE như sau:

WOE = ln(% Good / % Bad)

Trong đó:

* % Good: tỷ lệ mẫu "tốt" (ví dụ như trả nợ đúng hạn) trong một nhóm
* % Bad: tỷ lệ mẫu "xấu" (ví dụ như trả nợ không đúng hạn) trong một nhóm
* WOE càng lớn, biến số đó càng có khả năng phân biệt giữa nhóm "tốt" và nhóm "xấu".
  + - 1. Information Value (IV)

IV là một chỉ số đánh giá sức mạnh phân biệt của một biến số. Nó được tính dựa trên WOE của từng giá trị của biến số, theo công thức:

IV = Σ (% Good - % Bad) \* WOE

Trong đó:

* % Good: tỷ lệ mẫu "tốt" trong một nhóm
* % Bad: tỷ lệ mẫu "xấu" trong một nhóm
* WOE: Weight of Evidence của nhóm đó

Theo thông lệ, IV < 0.1 được coi là biến số yếu, 0.1 < IV < 0.3 là biến số trung bình, và IV > 0.3 là biến số mạnh. IV càng lớn, biến số đó càng có khả năng phân biệt giữa nhóm "tốt" và nhóm "xấu".

Việc tính toán và sử dụng WOE và IV giúp chúng ta lựa chọn các biến số quan trọng nhất để xây dựng mô hình scorecard dự báo rủi ro tín dụng một cách hiệu quả.

* 1. Tổng quan các nghiên cứu trước đây
     1. Mô hình chấm điểm tín dụng dựa trên rủi ro

Có rất nhiều nghiên cứu trước đây hay thậm chí là ứng dụng thực tế tại các tổ chức tài chính, người ta sử dụng nhiều mô hình học máy trong phân tích, nhưng được nghiên cứu nhiều nhất và đã ứng dụng vào thực tế là mô hình Hồi quy Logistic trong việc xây dựng một mô hình chấm điểm tín dụng dựa trên nền tảng rủi ro mà khách hàng mang lại.

1. Kablan, A. Yousfi, A Comparative Study of Machine Learning Algorithms for Credit Risk Assessment. Nghiên cứu này so sánh hiệu suất của các thuật toán học máy như Logistic Regression, Decision Trees, Random Forest và Gradient Boosting Machine để đánh giá rủi ro tín dụng. Nghiên cứu này tập trung vào so sánh và đánh giá hiệu suất của các thuật toán học máy phổ biến trong việc đánh giá rủi ro tín dụng. Kết quả cho thấy Random Forest và Gradient Boosting Machine thường cho kết quả tốt nhất trong việc dự đoán rủi ro tín dụng.
2. M. U. Hasan, M. S. Rahman, M. J. Alam tổng kết và phân tích các phương pháp học máy được sử dụng trong đánh giá rủi ro tín dụng. Tác giả thảo luận về ưu điểm và hạn chế của mỗi phương pháp và đề xuất hướng phát triển trong tương lai để cải thiện hiệu suất của các mô hình.

Lý do phổ biến mà các tổ chức hoặc các nhà nghiên cứu áp dụng thuật toán Hồi quy Logistic cho một quy trình phức tạp khi chấm điểm là vì tính giải thích của chúng với kết quả mà chúng đưa ra.

Sau khi xây dựng mô hình, các ngân hàng triển khai chấm điểm với quy trình như sau:



**Hình 1: Nhóm trường thông tin khách hàng**

A screenshot of a computer

Description automatically generated

Dữ liệu của từng khách hàng sẽ được phân theo từng nhóm trường thông tin. Sau đó khách hàng được chấm điểm ở mỗi trường dữ liệu này. Cuối cùng, điểm được tính tổng dựa trên trọng số đã quy định trước đó của chính trường dữ liệu. Trọng số zcó thể khác biệt ở các Ngân hàng khác nhau.

**Bước 1: Thu thập dữ liệu đầu vào**

Để tính điểm tín dụng khách hàng, Ngân hàng thu thập thông tin khách hàng theo từng nhóm dữ liệu đầu vào như sau

* **Dữ liệu đánh giá về thân nhân: T**ập trung vào các thông tin như tuổi, trình độ học vấn, lý lịch tư pháp, tình trạng hôn nhân, tình trạng cư trú, số người phụ thuộc, tình strạng sức khỏe của khách hàng. Trong số này, có 3 tiêu chí đánh giá được xác định tính chất, sự đánh giá phụ thuộc vào nhận định của cán bộ đánh giá như: Tình trạng sức khỏe của khách hàng, Đánh giá gia cảnh của khách hàng so với mặt bằng của vùng, đánh giá mối quan hệ của khách hàng so với cộng đồng.
* **Dữ liệu đánh giá khả năng trả nợ:** Thông thường nhóm này mô tả mức độ ổn định và khối lượng của nguồn thu nhập của khách hàng. Trong trường hợp khách hàng có thu nhập từ nhiều nguồn khác nhau, ngân hàng thường cân nhắc trên tổng nguồn thu nhập để đánh giá. Để đánh giá khả năng trả nợ của khách hàng, có thể chia thành hai phần: đánh giá mức độ ổn định của thu nhập (chiếm 47% điểm số của phần 2) và đánh giá nguồn thu nhập (chiếm 53% điểm số của phần 2).
* **Dữ liệu quan hệ tín dụng với Ngân hàng và các tổ chức tín dụng khác:** Bao gồm 5 tiêu chí để đánh giá mức độ sử dụng dịch vụ ngân hàng của khách hàng tại nội ngân hàng và tại các tổ chức tín dụng khác, đồng thời đánh giá chất lượng các khoản tín dụng của khách hàng tại các tổ chức tín dụng đó thông qua việc đánh giá lịch sử tín dụng và thiện chí trả nợ của khách hàng.

**Bước 2: Phân tích về trọng số:**

Mỗi ngân hàng sẽ có quy định riêng lẽ về trọng số cho từng khoản mục. Cấp một là trọng số theo (1) nhóm dữ liệu về nhân thân, (2) nhóm dữ liệu đánh giá khả năng trả nợ của khách hàng, (3) nhóm dữ liệu về quan hệ tín dụng.

**Bước 3: Tính tổng điểm:**

Sau khi thực hiện đánh giá ở trường nhóm trường dữ liệu, mô hình tính tổng điểm của khách hàng với công thức sau:

Trong đó:

z : Điểm tín dụng của khách hàng

: Điểm số của chỉ tiêu cấp 2 thứ i

: Hệ số rủi ro của chỉ tiêu cấp 2 thứ i

: Hệ số rủi ro của chỉ tiêu cấp 1 thứ j mà chỉ tiêu cấp 2 thứ i thuộc vào nó

**Bước 4: Xếp loại điểm số:**

Sau khi tính được tổng điểm của khách hàng, Ngân hàng xếp loại khách hàng và tiến hành các bước tìm điểm cutoff point phù hợp để chọn ra danh mục tốt nhất phù hợp với nguồn tài nguyên hiện hữu (giới hạn về dòng tiền, trần tín dụng,…)

* + 1. Mô hình chấm điểm tín dụng dựa trên lợi nhuận

Các nghiên cứu về việc chấm điểm tín dụng dựa trên lợi nhuận mặc dù khá mới trong những năm trở lại đây nhưng vẫn được rất nhiều nhà nghiên cứu trước đó đã triển khai thực hiện.

* Cụ thể, Tạp chí Banking and Finance số 2, 8.2019 đăng tải một bài viết tiêu đề “Profit scoring for credit unions using the multilayer perceptron, XGBoost and TabNet algorithms: Evidence from Peru” của nhóm tác giả Rodrigo Asencios, Christian Asencios, Efrain Ramos. Nghiên cứu này giới thiệu một phương pháp mới để đánh giá rủi ro tín dụng trong lĩnh vực tài chính, tập trung vào lợi nhuận thay vì chỉ sử dụng các tiêu chí truyền thống. Phương pháp này mang lại cái nhìn mới mẻ và hiệu quả về việc đánh giá và quản lý rủi ro tín dụng. Trong đó, bài nghiên cứu áp dụng một mô hình kết hợp của Multilayer Perceptron và XGBoost để xác định rủi ro tín dụng dựa trên góc nhìn lợi nhuận.
* Tạp chí Electronic Commerce Research and Applications, 2018 đăng tải một bài viết "Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score" tác giả Xin Ye, Lu-an Dong, Da Ma đã sử dụng một kết hợp của các thuật toán học máy, bao gồm Random Forest, Gradient Boosting, và Neural Networks
* Tạp chí Finance and Risk Management, 2023 đăng tải bài viết “Machine Learning-Based Profit Modeling for Credit Card Underwriting - Implications for Credit Risk” của tác giả George Krivorotov (một nhà kinh tế tài chính cao cấp Phòng Tính Toán và Phân Tích Rủi Ro của Văn Phòng Kiểm Toán của Cơ Quan Điều Hành Tiền Tệ (OCC)). Bài nghiên cứu so sánh hai mô hình máy học là XGBoost tính điểm lợi nhuận và Logistic regression tính điểm rủi ro của một khách hàng cá nhân đối với một tổ chức tín dụng. Trong bài viết, tác giả xây dựng mô hình mới sử dụng mô hình XGBoost (được phát triển năm 2018) với biến phụ thuộc là lợi nhuận từ phí và lãi, không bao gồm dự phòng. Trong bài viết tác giả cũng thành công tìm ra phân nhóm khách hàng có khả năng mang lại lợi nhuận cao đồng thời có rủi ro tương đối thấp. Tuy nhiên, tác giả cũng nêu ra quan ngại về tính chính xác trong việc dự đoán của mô hình tính điểm lợi nhuận do sự phức tạp trong hoạt động chi tiêu của khách hàng, cũng như cơ cấu thu nhập và chi phí của sản phẩm thẻ tín dụng.

CHƯƠNG 3: MÔ HÌNH TÍNH ĐIỂM LỢI NHUẬN

1. 1. Mô tả dữ liệu đầu vào

#### **Biến đầu vào**

Hình 3 dưới cho thấy dư nợ trung bình của thẻ giảm dần theo thời gian. Thông thường, thẻ được theo dõi trong 5 năm vì sau 5 năm dư nợ của thẻ quá nhỏ dẫn đến Lợi nhuận trong năm gần như không đóng góp vào Lợi nhuận trên tổng vòng đời khách hàng. Tuy nhiên, do giới hạn về mặt dữ liệu, phân tích sau sẽ dựa vào tập khách hàng trong năm 2021 và theo dõi trong vòng 2 năm đầu (tức 2 năm có khối lượng dư nợ tín dụng lớn nhất) trong vòng đời khách hàng. Ví dụ dễ hiểu là khách hàng A được phê duyệt cấp thẻ tín dụng vào tháng 01.2021 thì sẽ được theo dõi trong 2 năm đầu liên tục tức là đến cuối tháng 01.2023 và khách hàng B được phê duyệt cấp thẻ tín dụng vào tháng 12.2021 thì sẽ được theo dõi liên đến cuối tháng 12.2023

**Hình 3: Dư nợ trung bình của một khách hàng thay đổi qua 5 năm**

Dữ liệu sử dụng bao gồm 200.000 khách hàng cá nhân có sử dung thẻ tín dụng của cùng một ngân hàng trong nước được quan sát liên tục trong 24 tháng. Đặc điểm chung của tập dữ liệu mẫu là người Việt Nam trong độ tuổi lao động, thu nhập tháng trên 8 triệu tại thời điểm chấm điểm khách hàng. Trong đó bao gồm 49 trường dữ liệu về 3 nhóm dữ liệu nhân thân, khả năng trả nợ và quan hệ tín dụng từ các nguồn bao gồm dữ liệu từ nội bộ ngân hàng, từ CIC, và từ các đối tác được ghi nhận theo cấp độ cá nhân.

Bảng 2: Mô tả chi tiết từng nhóm dữ liệu:

|  |  |  |
| --- | --- | --- |
| **Nhóm thông tin nhân thân** | * Thông tin cơ bản: Tên , địa chỉ, id,… * Mô tả khách hàng: Công việc, thu nhập, tỷ lệ nợ/ thu nhập, mô tả khoản vay, cư trú,… * Phân loại phân khúc khách hàng định danh từ nhiều nguồn * Loại thẻ mà khách hàng muốn mở * Khách hàng mở thẻ qua kênh bán và chương trình gì | 28 biến |
| **Nhóm thông tin khả năng trả nợ** | * Thu nhập tự khai báo của khách hàng * Khối lượng ABD trong tài khoản Deposit | 2 biến |
| **Nhóm quan hệ tín dụng** | * Số lượng sản phẩm các loại hiện hữu tại ngân hàng * Số DPD cao nhất khách hàng có tại ngân hàng * Thuộc nhóm nợ trong vòng 36 tháng gần nhất * Thuộc nhóm nợ trong vòng 24 tháng gần nhất * Thuộc nhóm nợ trong vòng 12 tháng gần nhất * Thuộc nhóm nợ trong vòng 6 tháng gần nhất * Số lượng khoản vay tại ngân hàng * Hạn mức thẻ được phê duyệt, hạn mức khách hàng. (Tính toán trong một đề tài khác) * Nhóm CIC | 19 biến |

**Xử lý dữ liệu trống**: Hầu hết các trường dữ liệu đầy đủ, đối với dòng dữ liệu trống số sẽ được thay thế bởi trung vị của trường dữ liệu (median). Đối với dữ liệu chuỗi thì không có dòng dữ liệu trống.

Mặc dù dữ liệu đầu vào của hai mô hình được truyền vào mức độ từng khách hàng, tuy nhiên để mô tả danh mục khách hàng, bài viết sẽ phân tích thành 3 nhóm:

* Khách hàng không chi tiêu (Non-spender): Nhóm khách hàng chưa từng sử dụng thẻ trong trong khoản thời gian phân tích
* Khách hàng chi tiêu (Spending): Đây là nhóm khách hàng có phát sinh chi tiêu nhưng thanh toán toàn bộ số tiền đã sử dụng trong thẻ tín dụng sau ngày chốt sao kê và trước ngày tính lãi (due date). Họ không chịu lãi suất từ việc vay mượn tiền từ ngân hàng, mà chỉ tận dụng thẻ để chậm trả các khoản thực chi.
* Khách hàng dư nợ (Revolver): Có thể được chia tiếp thành 2 nhóm: (1) Nhóm A. Thanh toán một phần (vượt mức tối thiểu) và để lại số tiền lớn như là dư nợ. Nhóm này đóng lãi cho ngân hàng. Không bị ảnh hưởng bậc tín dụng CIC; (2) Nhóm B. Thanh toán dưới mức tối thiểu hoặc không thanh toán. Nhóm khách hàng vừa đóng lãi, vừa đóng phí phạt và bị đưa vào nhóm nợ cao hơn. Đồng thời, ngân hàng trích lập dự phòng theo nhóm cho nhóm B

Trong tập dữ liệu huấn luyện đầu vào, dữ liệu được ghi nhận trên cở sở Revolved rate (Tỷ suất dư nợ tính lãi). Trong đó: Revolving\_rate = số dư sinh lãi/số dư cuối kỳ. Cụ thể:

* Nhóm non-spending: Không có giao dịch chi tiêu trong thời gian quan sát
* Nhóm spender: có giao dịch chi tiêu và Revolved\_rate = 0
* Nhóm revolver: có giao dịch chi tiêu và Revolved\_rate ≠ 0

Việc chia theo nhóm trên có liên quan đến cơ cấu thu nhập theo vòng đời của khách hàng. Cụ thể, Hình 1 cho thấy với mọi thẻ chưa có chi tiêu (spending) thì chưa từng có nguồn thu nhập nào phát sinh trong thời gian quan sát, tuy nhiên trước và trong khi thẩm định thẻ cho khách hàng, các chi phí liên quan đến vận hành, mở thẻ,… đã xảy ra liên tục. Nói cách khác, nhóm khách hàng non-spending mới là đối tượng không nên được phê duyệt nhất nếu đứng từ góc độ lợi nhuận.

Đối với nhóm khách hàng spender, thu nhập từ nhóm này tập trung chủ yếu vào interchange, MCBS phát sinh do các giao dịch mà khách hàng đã thực hiện.

Đối với nhóm khách hàng revolver, thu nhập của nhóm này đa dạng vừa phát sinh từ các giao dịch khách hàng thực hiện, vừa phát sinh từ lãi và là nhóm ngân hàng đặc biệt quan tâm.

Hình 3: Chia bin dữ liệu

A computer screen shot of text

Description automatically generated

A screenshot of a graph

Description automatically generated

Hình 4: WOE và IV để tìm ra các biến và bin có ảnh hưởng lớn đến việc dự báo

A screenshot of a computer screen

Description automatically generated

A graph on a white background

Description automatically generated

Hình 4 phía trên đánh giá về WOE (Weight of Evidence) và IV (Information Value) cho đặc trưng 'age':

WOE (Weight of Evidence):

* Các giá trị WOE thể hiện logarit của tỷ lệ giữa tỷ lệ 'tốt' và tỷ lệ 'xấu' cho mỗi khoảng tuổi.
* Các giá trị WOE nằm trong khoảng từ -0.229418 cho khoảng tuổi đầu tiên ((-999999999999999.0, 27.0]) đến 0.343035 cho khoảng tuổi cuối cùng ([43.0, 999999999999999.0]).
* Các giá trị WOE dương cho thấy các khoảng tuổi cao hơn liên quan đến tỷ lệ 'tốt' cao hơn, trong khi các giá trị WOE âm gợi ý rằng các khoảng tuổi thấp hơn liên quan đến tỷ lệ 'xấu' cao hơn.

IV (Information Value):

* Các giá trị IV đo lường khả năng dự đoán của mỗi khoảng tuổi, với các giá trị cao hơn chỉ ra tầm quan trọng lớn hơn.
* Các giá trị IV nằm trong khoảng từ 0.000099 cho khoảng ((32.0, 34.0]) đến 0.018902 cho khoảng cuối cùng ([43.0, 999999999999999.0]).
* Các giá trị IV gợi ý rằng khoảng tuổi cuối cùng ([43.0, 999999999999999.0]) có khả năng dự đoán cao nhất, trong khi khoảng ((32.0, 34.0]) có khả năng dự đoán thấp nhất trong số các khoảng tuổi được cung cấp.

Nói chung, các giá trị WOE và IV cung cấp thông tin về mối quan hệ giữa tuổi và biến mục tiêu. Các giá trị WOE dương cho các khoảng tuổi cao hơn cho thấy những người lớn tuổi hơn liên quan đến tỷ lệ 'tốt' cao hơn, trong khi các giá trị WOE âm cho các khoảng tuổi thấp hơn gợi ý rằng những người trẻ tuổi hơn liên quan đến tỷ lệ 'xấu' cao hơn. Các giá trị IV làm nổi bật các mức độ khác nhau về khả năng dự đoán của mỗi khoảng tuổi, với khoảng cuối cùng ([43.0, 999999999999999.0]) là đặc trưng có thông tin hữu ích nhất.

Tương tự, các giá trị IV của từng trường dữ liệu như sau:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

#### **Biến đặc trưng**

##### ***Kỹ thuật biến đặc trưng***

Kỹ thuật biến đặc trưng (Feature engineering) là quá trình tạo ra các biến đặc trưng mới hoặc biến đổi các biến đặc trưng hiện có để cải thiện hiệu suất của mô hình. Mục tiêu của feature engineering là tạo ra các biến đặc trưng mang thông tin quan trọng và có khả năng phân biệt tốt giữa các trường hợp khác nhau. Các công việc trong feature engineering có thể bao gồm:

* Xử lý dữ liệu bị khuyết: Điền giá trị bị khuyết (missing values) hoặc loại bỏ các mẫu có dữ liệu bị khuyết.
* Rời rạc hóa (binning): Chuyển đổi biến liên tục thành biến rời rạc để tạo ra thông tin tương đồng và giảm nhiễu.
* Mở rộng đặc trưng (feature expansion): Tạo ra các biến đặc trưng mới từ các biến đặc trưng hiện có bằng cách sử dụng các phép toán, hàm số hoặc kết hợp các biến.
* Trích xuất đặc trưng (feature extraction): Trích xuất thông tin quan trọng từ dữ liệu gốc bằng cách sử dụng các phương pháp như phân tích thành phần chính (PCA) hay phân tích yếu tố (Factor Analysis).

##### ***Điểm quan trọng của biến đặc trưng***

Điểm quan trọng của biến đặc trưng (Feture importance) giúp chọn các biến đầu vào quan trọng đúng mức đối với dự đoán lợi nhuận của mô hình. Như chúng ta đã biết, chấm điểm tín dụng là một phần quan trọng trong quy trình cho vay, giúp các tổ chức tài chính quản lý rủi ro. Việc lựa chọn biến đặc trưng, một phần quan trọng trong quá trình đánh giá tính khả dụng tín dụng, thường nhằm mục tiêu làm giảm số lượng biến, từ đó giảm độ phức tạp của mô hình, chi phí thu thập dữ liệu và thời gian tính toán.

Với hai mô hình xây dựng bằng cách tiếp cận trên cơ sở lợi nhuận, tôi đã sử dụng thư viện được xây dựng sẵn cho Random Forest và XGBoost trong Python để đánh giá tầm quan trọng của các đặc trưng. Phương pháp này dựa trên cách mỗi đặc trưng được sử dụng để phân chia dữ liệu trong các cây quyết định của mô hình ensemble trees của XGBoost.

Hình 6.1: Feature Importance Score – XGboost

A graph with numbers and a number of text

Description automatically generated with medium confidence

Với Mô hình XGBoost, các biến đầu vào như hạn mức thẻ, độ tuổi khách hàng, học vấn là yếu tố ảnh hưởng nhiều nhất đến kết quả của sự dự đoán NFI. Tương tự, độ tuổi khách hàng, hạng mức vẫn là nhóm ảnh hưởng mạnh đến kết quả dự đoán trong NII.

Tuy nhiên, với mô hình Random Forest, các biến đầu vào như hạn mức thẻ, độ tuổi khách hàng, học vấn là yếu tố ảnh hưởng nhiều nhất đến kết quả của sự dự đoán NFI. Tương tự, độ tuổi khách hàng, hạng mức vẫn là nhóm ảnh hưởng mạnh đến kết quả dự đoán trong NII.

#### **Phân tích cơ cấu doanh thu thẻ tín dụng**

Phân tích mối tương quang Lợi nhuận thực tế và rủi ro dự đoán. Đối với cấu phần doanh thu của sản phẩm thẻ tín dụng có thể được chia ra thành 2 cấu phần chính là Lãi nằm trong NII và Phí nằm trong NFI. Chi phí và tổn thất phát sinh trong thẻ tín dụng cũng là các yếu tố quan trọng của tính lợi nhuận từ thẻ. Trong đó các khoản phí (fee) bao gồm các khoản phí đa dạng như phí rút tiền mặt, phí trễ hạn, hoặc phí hội viên hàng năm. Các khoản phí giao dịch (interchange) thay đổi dựa trên các loại giao dịch, chẳng hạn như giữa các khoản phí du lịch và chi tiêu sinh hoạt tại các siêu thị.

##### **Cấu phần NII**

NII (Net Interest Income) hay Thu nhập lãi thuần là một thành phần quan trọng trong hoạt động thẻ tín dụng của ngân hàng. Để phân tích cấu phần tạo ra NII trong hoạt động thẻ tín dụng, cần xem xét các yếu tố chính: Lãi thu từ dư nợ thẻ tín dụng, Chi phí lãi trả cho nguồn vốn, Tỷ lệ nợ xấu và các khoản phải trích lập dự phòng

* **Lãi thu từ dư nợ thẻ tín dụng**

**Lãi suất áp dụng**

* Lãi suất thường niên: Đây là lãi suất mà ngân hàng áp dụng cho số dư nợ của khách hàng nếu không thanh toán đầy đủ số dư vào cuối kỳ sao kê. Lãi suất này thường cao hơn lãi suất vay thông thường do rủi ro tín dụng cao hơn.
* Phí trễ hạn: Ngoài lãi suất, ngân hàng có thể thu thêm các khoản phí nếu khách hàng không thanh toán đúng hạn. Những khoản phí này cũng góp phần vào tổng lãi thu.

**Số dư nợ thẻ tín dụng**

* Tổng dư nợ: Tổng dư nợ chưa thanh toán của khách hàng càng lớn, thu nhập lãi càng cao.
* Mức sử dụng thẻ: Mức độ sử dụng thẻ của khách hàng (số lượng giao dịch và giá trị giao dịch) ảnh hưởng trực tiếp đến dư nợ.
* **Chi phí lãi trả cho nguồn vốn**

**Chi phí huy động vốn**

* Lãi suất huy động: Đây là lãi suất ngân hàng phải trả cho các khoản tiền gửi và các nguồn vốn khác mà ngân hàng sử dụng để cấp tín dụng cho khách hàng. Sự chênh lệch giữa lãi suất huy động và lãi suất cho vay là yếu tố chính tạo ra NII.
* Cơ cấu nguồn vốn: Ngân hàng sử dụng các nguồn vốn khác nhau với các chi phí huy động khác nhau. Cơ cấu này ảnh hưởng đến tổng chi phí lãi phải trả.

**Tỷ lệ duy trì nguồn vốn**

* Dự trữ bắt buộc: Ngân hàng phải duy trì một tỷ lệ nhất định dự trữ bắt buộc tại ngân hàng trung ương, không sinh lãi hoặc lãi suất thấp, làm tăng chi phí cơ hội.
* Quản lý thanh khoản: Đảm bảo đủ thanh khoản để đáp ứng các yêu cầu rút tiền và thanh toán từ khách hàng cũng ảnh hưởng đến chi phí huy động.
* **Tỷ lệ nợ xấu và các khoản phải trích lập dự phòng**

**Nợ xấu**

* Tỷ lệ nợ xấu (NPL): Tỷ lệ nợ xấu càng cao, ngân hàng phải dành nhiều nguồn lực hơn để xử lý và trích lập dự phòng, làm giảm NII.
* Phục hồi nợ xấu: Khả năng thu hồi nợ xấu cũng ảnh hưởng đến thu nhập lãi thuần.

**Trích lập dự phòng rủi ro tín dụng**

* Chi phí dự phòng: Ngân hàng phải trích lập dự phòng cho các khoản nợ có khả năng không thu hồi được. Chi phí này giảm thu nhập lãi thuần.

##### **Cấu phần NFI**

* **Ngân hàng phát hành thẻ (Issuer)**

Khi ngân hàng A là ngân hàng phát hành thẻ, nó sẽ thu một số loại phí và cũng phải chịu một số chi phí nhất định liên quan đến giao dịch thẻ của khách hàng.

Các loại phí mà ngân hàng A nhận được:

* Phí liên ngân hàng (Interchange Fee): Khi khách hàng của ngân hàng A sử dụng thẻ để mua hàng từ một thương nhân, ngân hàng A sẽ nhận được phí liên ngân hàng từ ngân hàng của thương nhân. Mức phí này phụ thuộc vào MCC của thương nhân, loại thẻ, và loại giao dịch. Thu nhập từ phí interchange là một loại thu nhập mà các tổ chức tài chính như ngân hàng hoặc tổ chức thẻ tín dụng nhận được khi các giao dịch thanh toán bằng thẻ được thực hiện. Phí interchange là khoản phí mà người bán (merchant) phải trả cho tổ chức tài chính của họ mỗi khi họ chấp nhận một giao dịch thanh toán bằng thẻ từ khách hàng. Tổ chức tài chính đó sau đó chia khoản phí này với tổ chức phát hành thẻ (ví dụ: ngân hàng phát hành thẻ tín dụng của người tiêu dùng). Ví dụ, giả sử khách hàng A đi mua một cốc cà phê tại một cửa hàng cà phê và thanh toán bằng thẻ tín dụng Visa ở Ngân hàng B. Sau khi giao dịch được thực hiện, cửa hàng cà phê phải trả một khoản phí nhất định cho Ngân hàng C sở hữu máy Pos, được gọi là phí interchange. Số tiền này có thể là một phần trăm nhỏ của tổng giá trị giao dịch. Phần của khoản phí này sau đó sẽ được chia giữa Ngân hàng B và Ngân hàng C.

Các loại chi phí mà ngân hàng A phải trả:

* Phí mạng lưới (Network Fees): Ngân hàng A phải trả phí mạng lưới cho các tổ chức như Visa và MasterCard để sử dụng mạng lưới thanh toán của họ. Phí này có thể bao gồm phí xử lý giao dịch và phí duy trì mạng lưới.
* Phí bảo mật và tuân thủ: Ngân hàng A cũng có thể phải trả các phí liên quan đến bảo mật giao dịch và tuân thủ các quy định của mạng lưới thẻ.
* **Ngân hàng của thương nhân (Acquirer)**

Khi ngân hàng A là ngân hàng của thương nhân, nó xử lý các giao dịch thẻ cho các thương nhân.

Các loại phí mà ngân hàng A nhận được:

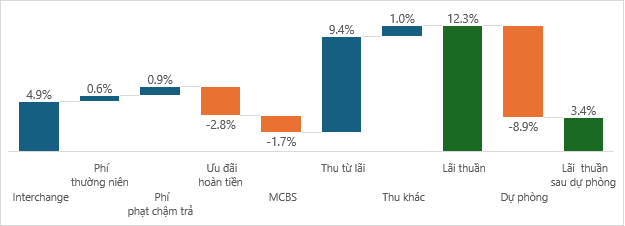
* Phí thương nhân (Merchant Discount Rate - MDR): Ngân hàng A thu phí từ các thương nhân dựa trên giá trị giao dịch. Mức phí này thường được cấu thành từ phí liên ngân hàng, phí mạng lưới, và một phần phí dịch vụ mà ngân hàng A thu.
* Phí dịch vụ: Ngân hàng A có thể thu thêm phí dịch vụ từ thương nhân cho việc cung cấp thiết bị POS, dịch vụ hỗ trợ và các dịch vụ giá trị gia tăng khác.

Các loại chi phí mà ngân hàng A phải trả:

* Phí liên ngân hàng (Interchange Fee): Ngân hàng A phải trả phí liên ngân hàng cho ngân hàng phát hành thẻ của khách hàng. Mức phí này phụ thuộc vào MCC của thương nhân và loại giao dịch.
* Phí mạng lưới (Network Fees): Ngân hàng A phải trả phí mạng lưới cho Visa, MasterCard và các tổ chức thanh toán khác. Phí này bao gồm phí xử lý giao dịch và phí duy trì mạng lưới.
* Phí tuân thủ và bảo mật: Ngân hàng A phải chi trả các chi phí liên quan đến tuân thủ quy định và bảo mật để bảo vệ giao dịch và dữ liệu khách hàng.

Trong tập dữ liệu, tỷ trọng trung bình giữa các dòng thẻ tín dụng đóng góp vào thu nhập thẻ có thể dược mô tả như sau:

**Hình 4: Cấu phần tài chính bình quân năm của thẻ trên Dư nợ trung bình**



Hình 3 phân tích về cấu phần các nguồn doanh thu chi phí trọng yếu của thẻ tín dụng. Với 1 đồng cho vay khách hàng, trung bình ngân hàng thu được 0.12 đồng so với dư nợ trung bình. Tỷ trọng đóng góp quan trọng nhất vẫn là NII đóng góp bởi nhóm khách hàng có dư nợ phát sinh lãi là 13% và phí interchange đóng góp bởi hoạt động chi tiêu của khách hàng. Song song đó cũng có 3 nhóm chi phí chính là Hoàn tiền khách hàng dựa trên chi tiêu trong tháng (4%) và MCBS (4%). Ngoài ra, Provision cũng chiếm tỷ trọng lớn trong nhóm làm giảm thu nhập ngân hàng (4%).

**Hình 5: Cấu phần doanh thu và chi phí**

* 1. Xây dựng mô hình

#### **Chọn mô hình machine learning**

Trong mô hình sau, biến độc lập được truyền vào được mô tả ở Bảng 2, và biến phụ thuộc được sử dụng là thu nhập lãi thuần (TOI). Thu nhập TOI được tạo ra từ hai cấu phần là Thu nhập sau chi phí từ lãi (NII) và Thu nhập sau chi phí từ phí (NFI). Từ đó, khi thực hiện dự đoán, kết quả được đưa ra là các biến liên tục dự đoán về thu nhập mà khách hàng đó có thể mang lại cho ngân hàng.

Về mặt lý thuyết, các mô hình phổ biến đã từng được nghiên cứu và áp dụng trong bài toán về xây dựng mô hình điểm tín dụng đều có ưu và nhược điểm riêng. Cụ thể:

Logistic Regression

* (1) Đơn giản và dễ giải thích: Logistic Regression là một trong những mô hình đơn giản nhất và dễ hiểu nhất. Điều này rất quan trọng vì chúng giúp kết quả trở nên minh bạch hơn vì khả năng giải thích của mô hình là yếu tố then chốt trong quản trị rủi ro. (2) Hiệu quả với dữ liệu tuyến tính: Logistic Regression hoạt động rất tốt với các dữ liệu có mối quan hệ tuyến tính giữa các biến đầu vào và biến đầu ra. (3) Tiết kiệm tài nguyên máy: Mô hình Logistic Regression yêu cầu ít tài nguyên tính toán hơn so với các mô hình phức tạp hơn, cho phép huấn luyện nhanh và triển khai dễ dàng.
* Tuy nhiên, (1) Hiệu suất kém với dữ liệu phi tuyến tính: Logistic Regression giả định mối quan hệ tuyến tính giữa các biến đầu vào và đầu ra. Với dữ liệu phức tạp và phi tuyến tính, mô hình này có thể không thể nắm bắt được các mẫu quan trọng. (2) Đòi hỏi các biến độc lập: Logistic Regression giả định rằng các biến đầu vào độc lập với nhau. Nếu có sự tương quan cao giữa các biến, mô hình có thể gặp khó khăn. (3) Overfitting hoặc Underfitting: Với dataset nhỏ, Logistic Regression có thể dễ dàng bị overfitting nếu số lượng biến độc lập lớn so với số lượng mẫu, hoặc underfitting nếu mô hình quá đơn giản để nắm bắt các đặc trưng của dữ liệu.
* Do đó, đòi hỏi các hoạt động xử lý dữ liệu nâng cao như (1) Regularization: Sử dụng kỹ thuật regularization (như L1 hoặc L2) để giảm overfitting. (2) Feature Engineering: Tạo ra các đặc trưng mới hoặc sử dụng kỹ thuật giảm chiều (dimensionality reduction) để giảm thiểu tác động của các biến tương quan.

Random Forest

* (1) Hiệu suất Cao: Random Forest thường đạt hiệu suất cao trên nhiều loại dữ liệu khác nhau, bao gồm cả dữ liệu phi tuyến tính và không cân bằng. (2) Giảm Thiểu Overfitting: Bằng cách kết hợp nhiều cây quyết định (decision trees) và sử dụng kỹ thuật bagging, Random Forest giảm thiểu vấn đề overfitting. (3) Tính Ổn Định và Khả Năng Mở Rộng: Random Forest cung cấp kết quả ổn định và có thể dễ dàng mở rộng cho các tập dữ liệu lớn.
* Tuy nhiên, (1) Overfitting: Mặc dù Random Forest giảm overfitting so với một cây quyết định đơn lẻ, nhưng với dataset nhỏ, nó vẫn có thể bị overfitting, đặc biệt khi số lượng cây lớn và các cây không được giảm bớt đủ. (2) Thiếu tính khái quát hóa: Random Forest có thể gặp khó khăn trong việc khái quát hóa từ một dataset nhỏ do mỗi cây chỉ nhìn thấy một phần nhỏ của dữ liệu.
* Do đó, đòi hỏi các hoạt động xử lý dữ liệu thô: Giảm số lượng cây: Sử dụng một số lượng cây vừa phải để tránh overfitting. Pruning: Áp dụng kỹ thuật pruning để giảm kích thước của các cây quyết định. Bootstrap samples: Sử dụng mẫu bootstrap với tỉ lệ thấp hơn để xây dựng từng cây quyết định.

XGBoost

* Hiệu suất Vượt Trội: GBM và đặc biệt là XGBoost được biết đến với hiệu suất vượt trội trên nhiều loại dữ liệu khác nhau. XGBoost là một cải tiến của GBM với nhiều tối ưu hóa về tốc độ và hiệu suất. Xử Lý Tốt Dữ Liệu Không Cân Bằng: Các mô hình này có khả năng xử lý tốt dữ liệu không cân bằng, điều thường gặp trong credit scoring, nơi số lượng khách hàng tốt thường lớn hơn số lượng khách hàng xấu. Khả Năng Điều Chỉnh Cao: GBM và XGBoost cung cấp nhiều tham số để điều chỉnh, cho phép tối ưu hóa hiệu suất mô hình. Khả Năng Giải Thích Tương Đối Tốt: Mặc dù phức tạp hơn Logistic Regression, các mô hình này vẫn có thể được giải thích ở mức độ tương đối thông qua các công cụ như SHAP (SHapley Additive exPlanations) và Feature Importance.
* Overfitting: GBM và XGBoost rất mạnh mẽ nhưng cũng rất dễ bị overfitting, đặc biệt khi dataset nhỏ và mô hình quá phức tạp. Thời gian huấn luyện dài: Với dataset nhỏ, thời gian huấn luyện có thể không quá dài, nhưng các mô hình GBM và XGBoost vẫn yêu cầu nhiều bước huấn luyện và điều chỉnh tham số, có thể trở nên quá mức cần thiết với tập dữ liệu nhỏ. Yêu cầu điều chỉnh tham số: GBM và XGBoost yêu cầu nhiều tham số cần được tối ưu hóa, và với dataset nhỏ, việc điều chỉnh tham số có thể khó khăn và dễ dẫn đến overfitting.
* Early Stopping: Sử dụng kỹ thuật early stopping để ngăn chặn mô hình huấn luyện quá mức. Regularization: Áp dụng các tham số regularization (như lambda và alpha trong XGBoost) để giảm overfitting. Cross-validation: Sử dụng k-fold cross-validation để đánh giá mô hình một cách chắc chắn hơn và giảm nguy cơ overfitting.

#### **Huấn luyện mô hình**

Sau nhiều nghiên cứu cân nhắc về điểm mạnh và điểm yếu của các mô hình khách nhau, mô hình cuối cùng được sử dụng là Extreme Gradient Boosting (XGBoost). Mặc dù hạn chế về tính giải thích trong các kết quả mà nó đưa ra và cũng dễ gặp tình trạng quá khớp nếu không được xử lý tốt, XGBoost vẫn là mô hình có khả năng cho ra kết quả gần với thực tế nhất.

Table 2: Siêu tham số cho từng mô hình con

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mô hình con** | **Gamma** | **Tốc độ học** | **Tầng cây** | **Số lượng cây** |
| **XGBoost** | | | | |
| NII | 0.01 | 0.1 | 7 | 100 |
| NFI | 0.01 | 0.2 | 3 | 100 |
|  | Số lượng nút | Số đặc trưng | Tầng cây | Số lượng cây |
| **Random Forest** | | | | |
| NII | 52 | 17 | 7 | 100 |
| NFI | 52 | 17 | 7 | 100 |

Trong đó:

* Tốc độ học: tham số điều chỉnh liên quan đến việc quyết định có nên chia một node trong cây hay không. Cụ thể, nó là ngưỡng tối thiểu mà một split (chia tách) mới phải vượt qua để được thực hiện
* Gamma: cao hơn sẽ làm cho mô hình ít phức tạp hơn, vì nó sẽ yêu cầu một sự cải thiện lớn hơn trước khi thực hiện chia tách. Điều này có thể giúp giảm hiện tượng overfitting (quá khớp)
* Tầng cây: Các giá trị khả dĩ cho độ sâu tối đa của cây là 3, 5 và 7. Độ sâu lớn hơn có thể giúp mô hình học được các mối quan hệ phức tạp hơn từ dữ liệu, nhưng cũng có nguy cơ dẫn đến quá khớp (overfitting)
* Số lượng cây: Số lượng cây trong rừng ngẫu nhiên. Chỉ định rằng mô hình sẽ sử dụng 100 cây. Số lượng cây càng nhiều thì mô hình có thể đạt được độ chính xác cao hơn, nhưng cũng làm tăng thời gian tính toán.
* Số lượng nút: Số lượng nút lá tối đa trong mỗi cây. Các giá trị khả dĩ cho số lượng nút lá tối đa là 50, 51 và 52. Giới hạn số lượng nút lá có thể giúp tránh quá khớp bằng cách kiểm soát độ phức tạp của cây
* Số đặc trưng: Số lượng đặc trưng tối đa được xem xét khi tìm kiếm sự phân tách tốt nhất. Các giá trị khả dĩ cho số lượng đặc trưng tối đa là 15, 16 và 17. Giá trị này điều chỉnh độ ngẫu nhiên của mỗi cây; số lượng đặc trưng ít hơn có thể làm giảm độ tương quan giữa các cây và tăng cường hiệu suất tổng thể

Sau khi thực hiện tìm kiếm siêu tham số (hyperparameter) với kỹ thuật tìm kiếm lưới (grid search) để điều chỉnh các siêu tham số độ sâu của tầng cây (tree depth) và tốc độ học (learning rate). Sử dụng kỹ thuật kiểm chứng chéo với số mẫu là 5 (cross-validation 5-fold). Tỷ lệ học (learning rate) ban đầu được thiết lập tìm trong khoảng từ 0.001 đến 0.01, do rủi ro cao từ việc quá khớp trên mẫu này, và do đó số lượng cây cũng cao. Tôi chọn tham số siêu tối ưu dựa trên Mean Squared Error (MSE) trên tập kiểm tra ngoài mẫu cho các mô hình con loại hồi quy và lỗi phân loại trên tập kiểm tra ngoài mẫu cho mô hình PD, sau đó tìm ra 30 biến hàng đầu cho mỗi mô hình con dựa trên chỉ số "gain" của XGBoost. Sau đó, tôi tiến hành ước lượng lại tất cả các mô hình con trên không gian đặc trưng hạn chế này với một tìm kiếm tham số siêu đầy đủ khác. Kết quả của các tìm kiếm này có thể được xem trong bảng.

* 1. Chỉ số đánh giá khả năng dự đoán

#### **Đánh giá mô hình XGboost**

Đánh giá về hiệu suất mô hình, với dữ liệu ở mục Tương quan rủi ro và Lợi nhuận trong thực nghiệm, ta có thể thấy mô hình dự đoán so với thực tế có khả năng dự đoán kém nhưng hiệu suất xếp hạng tốt hơn. Ở đây, khi dữ liệu về hành vi khách hàng không có đủ nhiều, việc dự đoán giá trị của khách hàng đối với ngân hàng là một thách thức lớn. Ví dụ, rõ ràng rằng mô hình gặp khó khăn trong việc ghi nhận sự biến đổi của lợi nhuận thực tế cho từng nhóm rủi ro và không thể ghi nhận được sự thay đổi trong lợi nhuận thực tế tại các nhóm khách hàng có rủi ro cao hơn. Điều này có thể ảnh hưởng đến tính tin cậy của dự đoán lợi nhuận trong các nhóm rủi ro cao, nơi mô hình dựa trên lợi nhuận mang tính thông tin nhất so với mô hình dựa trên rủi ro.

Với dự đoán giá tị NII, mô hình cho ra R2 10.9% và NFI là 18.8%. Tương tự với kết quả về khả năng dự báo trong mô hình và tập dự liệu của XGBoost mà George Krivorotov đã cho ra với Mô hình Profit (trong bài viết này là TOI) tại ngưỡng R2 ngưỡng 9 – 11% tùy tập khách hàng.

Như đã trình bày trước đó về đặc trưng của các mô hình học máy và học sâu, đặc trưng heuristics có thể dẫn đến tình trạng quá khớp từ đó đưa ra các dự đoán sai trong mẫu kiểm tra. Với phương pháp kiểm định chéo siêu tham số, XGBoost có thể giảm thiểu tính quá khớp. Tuy nhiên việc cải thiện khả năng giải thích của mô hình còn cần được phân tích sâu hơn.

Một khía cạnh của mô hình được xem xét là sự quan trọng của các đặc điểm giữa các danh mục tổng hợp và trên tổng thể. XGBoost bao gồm một phương pháp tính quan trọng của biến mặc định tính toán mức độ mỗi điểm chia cải thiện hiệu suất, được cân nhắc theo số lượng quan sát mà nút đó chịu trách nhiệm. Tuy nhiên, phương pháp này được thiết kế đặc biệt cho các mô hình cây tổ hợp và không thể tự động được áp dụng cho đầu ra tổng hợp NPV, mà là một sự kết hợp của nhiều mô hình cây tổ hợp. Thay vào đó, đối với mô hình tổng hợp, tôi sử dụng phương pháp quan trọng của biến từ gói iml R, với cách tiếp cận không phụ thuộc vào mô hình và ngẫu nhiên hóa các biến một cách tuần tự để xếp hạng theo sự mất đi trong khả năng dự đoán. Kết quả của phân tích này được hiển thị trong hình 8

#### **Đánh giá mô hình Random Forest**

Mô hình Random Forest cho kết quả Mean Squared Error (MSE) là 10.5 và R-squared (R²) là 85% và 89% lần lượt cho mô hình con NII và NFI. Điều này cho thấy mô hình có khả năng dự đoán tốt và giải thích được biến thiên của profit-score.

Hiệu Suất: Mô hình Random Forest cho thấy hiệu suất dự đoán tốt với R² cao, chứng tỏ mô hình có thể giải thích được phần lớn biến thiên của profit-score.

Tầm Quan Trọng của Đặc Trưng: Việc sử dụng Random Forest giúp xác định các đặc trưng quan trọng nhất ảnh hưởng đến profit-score, giúp ngân hàng tập trung vào các yếu tố này để tối ưu hóa lợi nhuận.

Ứng Dụng Thực Tế: Mô hình có thể được áp dụng trong thực tế để đánh giá điểm tín dụng của khách hàng, từ đó đưa ra các quyết định về cho vay và quản lý rủi ro hiệu quả hơn.

Mặc dù vậy, có thể thấy được rằng mô hình Random Forest có hạn chế như sau:

* Khả Năng Giải Thích: Mặc dù mô hình Random Forest có hiệu suất cao, nhưng khả năng giải thích của mô hình thấp hơn so với các mô hình đơn giản hơn như hồi quy tuyến tính.
* Tài Nguyên Tính Toán: Việc huấn luyện và dự đoán với mô hình Random Forest đòi hỏi tài nguyên tính toán lớn hơn, đặc biệt khi làm việc với các tập dữ liệu rất lớn.

#### **Kết luận trên hai mô hình tính điểm lợi nhuận**

Dựa trên kết quả trên, tôi thực lựa chọn mô hình XGBoost để triển khai với những giải thích như sau:

**Hiệu Suất Cao Hơn**

* XGBoost (Extreme Gradient Boosting) thường có hiệu suất cao hơn so với Random Forest nhờ vào các kỹ thuật tối ưu hóa như regularization, xử lý missing values, và khả năng chống overfitting tốt hơn.
* Overfitting Control: XGBoost có các tham số như gamma, lambda, và alpha giúp kiểm soát overfitting một cách hiệu quả. Điều này làm cho mô hình XGBoost có khả năng tổng quát hóa tốt hơn trên dữ liệu kiểm tra.
* Tree Pruning: XGBoost sử dụng kỹ thuật pruning giúp giảm độ sâu của cây quyết định, loại bỏ các nhánh không cần thiết và tăng cường tính tổng quát của mô hình.

**Xử Lý Dữ Liệu Missing Values Tốt Hơn**

XGBoost có khả năng xử lý missing values một cách tự động bằng cách tìm ra hướng phân chia tốt nhất, ngay cả khi dữ liệu bị thiếu. Điều này giúp cải thiện độ chính xác của mô hình mà không cần phải xử lý dữ liệu bị thiếu một cách thủ công.

**Tối Ưu Hóa Tốc Độ Huấn Luyện**

XGBoost được tối ưu hóa cho cả thời gian huấn luyện và sử dụng bộ nhớ hiệu quả hơn so với Random Forest nhờ vào các thuật toán song song và tăng cường hiệu suất tính toán.

Thời Gian Huấn Luyện: XGBoost sử dụng phương pháp tree boosting theo hướng chia nhỏ tập dữ liệu và huấn luyện song song, giúp giảm thời gian huấn luyện so với Random Forest.

**Khả Năng Giải Thích và Hiểu Biết Mô Hình**

* XGBoost cung cấp nhiều công cụ và chỉ số để giải thích mô hình, chẳng hạn như:
* Feature Importance: XGBoost có thể cung cấp tầm quan trọng của từng đặc trưng, giúp hiểu rõ hơn về những yếu tố nào ảnh hưởng lớn nhất đến profit-score.
* SHAP Values: Các giá trị SHAP (SHapley Additive exPlanations) giúp giải thích rõ ràng và trực quan sự ảnh hưởng của từng đặc trưng đối với dự đoán của mô hình.

**Độ Chính Xác Cao Hơn**

XGBoost cho kết quả độ chính xác cao hơn nhờ vào việc tối ưu hóa hàm mất mát và sử dụng thuật toán boosting để liên tục cải thiện mô hình dựa trên lỗi của các cây trước đó.

Do đó, việc chọn mô hình XGBoost để xây dựng scorecard profit có nhiều lợi thế hơn so với Random Forest, bao gồm hiệu suất cao hơn, khả năng xử lý missing values tốt hơn, thời gian huấn luyện ngắn hơn, và khả năng giải thích mô hình tốt hơn. Những yếu tố này giúp XGBoost trở thành lựa chọn ưu việt cho việc xây dựng scorecard profit trong lĩnh vực tài chính.

CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU

* 1. Kết quả mô hình

Để giải thích cho câu hỏi đặt ra trước đó về việc liệu thay đổi phương pháp và sử dụng mô hình Chấm điểm theo lợi nhuận thay cho Chấm điểm theo rủi ro thì sẽ tác động như thế nào đến danh mục khách hàng, tôi tiến hành lựa chọn tập khách hàng có lợi theo hai mô hình mà trong lĩnh vực thẩm định tín dụng gọi tắt là tìm điểm cutoff. Theo đó, mô hình đi tìm X% khách hàng tốt nhất với mô hình Chấm điểm theo lợi nhuận, thì tập X% khách hàng này sẽ cho ra khả năng vỡ nợ như thế nào so với tập X’% Khách hàng tốt nhất dựa trên mô hình Chấm điểm theo rủi ro với cùng một mức tổng tín dụng cấp cho danh sách khách hàng X%?

Sau khi đánh giá hai mô hình như kết luận trước đó, tôi lựa tiếp cận mô hình XGBoost và cho ra kết quả như sau:

Hình 7: Tương quan điểm rủi ro và điểm lợi nhuận

A graph of a graph with a blue line

Description automatically generated

Hình 7 cho thấy khi điểm tín dụng dựa trên rủi to tăng dần thì khả năng xảy ra vỡ nợ của khách hàng càng thấp. Tuy nhiên, đối với mô hình tính điểm lợi nhuận khách hàng thì khi điểm lợi nhuận càng tăng, xác suất vỡ nợ của khách hàng cũng tăng theo một cách rất chậm. Hơn hết, tại ngưỡng điểm cutoff tại 85-90, xác suất vỡ nợ của khách hàng bắt đầu giam, thậm chí tại ngưỡng điểm 90-95 trong mô hình điểm lợi nhuận thì tập khách hàng này bắt đầu có xác suất vỡ nợ thấp hơn nhóm từ dưới 10 điểm.

Điều này cho thấy nhóm khách hàng có điểm số trên 85 trong mô hình tính điểm lợi nhuận là tập khách hàng rất tốt, vừa được đánh giá là mang lại lợi nhuận rất tốt mà khả năng vỡ nợ của nhóm này cũng tương đối thấp.

* 1. Các điểm cần cải thiện

Với Học máy, dữ liệu đầu vào đóng một phần rất quan trọng đến kết quả đầu ra của mô hình. Tuy nhiên, vì giới hạn về dữ liệu nên mô hình trên chỉ được truyền vào 50 trường dữ liệu so với trung bình 100 trường dữ liệu dù các trường này có tính trùng lắp cao.

Một trong số những trường dữ liệu bị thiếu nhưng rất cần để hoàn thiện mô hình là Ddự phòng theo cấp độ mỗi khách hàng. Vì dự phòng về tổng thể chiếm tỷ trọng gần 9% ADB và là mục chi phí lớn nhất trước khi tính ra lợi nhuận sau cùng mà một khách hàng có thể mang lại (Hình 2).

Tuy nhiên, do đặc thù hoạt động trích lập dự phòng là theo kỳ, mức trích lập sẽ thay đổi theo thời gian, việc ghi nhận dự phòng có thể đúng trong thời điểm ghi nhận nhưng có thể thay đổi dưới góc độ suốt vòng đời khách hàng[[3]](#footnote-3). Do đó, việc ghi nhận Chi phí dự phòng trong khoản thời gian quan sát mang tính tương đối và nên được theo dõi liên tục trong nhiều năm hơn nữa khi thực tế ứng dụng mô hình vào hoạt động kinh doanh của Ngân hàng.

Trường hợp thành công lấy được dự phòng theo cấp độ khách hàng một cách chính xác, có nghĩa là mô hình có biến phụ thuộc là các chỉ số hậu Dự phòng (Ví dụ như PBT hoặc TOI – PRO) thì việc ra quyết định khách hàng sẽ đơn giản hơn vì khi này kết quả vừa phản ánh lợi nhuận vừa phản ánh rủi ro của khách hàng. Tuy nhiên, cùng với sự thành công trong việc xây dựng hệ thống công nghệ hạ tầng tại các ngân hàng trong nước, việc phân tách và tính được chỉ số này là một điều có khả năng xảy ra.

KẾT LUẬN

Kết quả của mô hình cho thấy, dựa trên một nhóm nhỏ khách hàng có điểm lợi nhuận từ trên 85 điểm, mô hình đưa ra kết quả tốt về mặt lợi nhuận và cũng chấp nhận được về mặt rủi ro tín dụng. Tuy nhiên, cần lưu ý rằng mô hình hiện tại có một số hạn chế do giới hạn về độ dày của lịch sử dữ liệu và độ rộng của các trường dữ liệu của khách hàng.

Một trong những hạn chế là mô hình hiện tại đạt được mức độ dự đoán khá thấp, được đo bằng chỉ số R2. Điều này có nghĩa là mô hình không thể giải thích một phần lớn sự biến động trong dữ liệu và có thể không thể dự đoán chính xác lợi nhuận cho tất cả các khách hàng. Việc này có thể xuất phát từ việc giới hạn dữ liệu lịch sử chỉ bao phủ một khoảng thời gian hạn chế hoặc không có đủ đa dạng thông tin về các khía cạnh quan trọng của khách hàng.

Tuy nhiên, với sự phát triển của hạ tầng dữ liệu trong nước, việc thu thập và xử lý dữ liệu đã trở nên ngày càng hoàn thiện hơn. Điều này cung cấp cơ hội để sử dụng bộ dữ liệu lớn hơn và đa dạng hơn, từ đó nâng cao khả năng dự đoán của mô hình. Điều quan trọng là cần tiến hành việc truyền đầu vào cho mô hình với một bộ dữ liệu phù hợp và đủ rộng, bao gồm các thông tin quan trọng về khách hàng như lịch sử tài chính, hành vi mua hàng và thông tin cá nhân.

Việc triển khai một cách thận trọng với mô hình tính điểm lợi nhuận dựa trên máy học là khả thi và có thể mang lại lợi ích đáng kể. Tuy mô hình hiện tại có mức độ dự đoán tương đối thấp, nhưng với sự cải thiện của dữ liệu và việc tinh chỉnh mô hình, có thể tăng khả năng dự đoán và cải thiện hiệu suất của mô hình. Tuy nhiên, cần đảm bảo rằng việc triển khai mô hình được thực hiện cẩn thận, bao gồm việc kiểm tra và đánh giá kỹ lưỡng để đảm bảo tính tin cậy và hiệu quả của mô hình trong thực tế.

Danh mục tài liệu tham khảo

1. George Krivorotov, 2023*. Machine Learning-Based Profit Modeling for Credit Card Underwriting - Implications for Credit Risk*
2. Nhóm tác giả của Google, 2024. *Hồi Quy Logistic*
3. Minghua Chen và Qun-Ying Liu, 2019. *XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System*
4. Leo Breima, 2001. *Random Forests Machine Learning* (1.2.5, Điểm hoán vị quan trọng)
5. Apley and Zhu, 2019. *Visualizing the effects of predictor variables in black-box supervised learning models* (2.2.4, phương pháp ALE)
6. Kevin P. Murphy, 2012. Machine Learning: A Probabilistic Perspective. MIT Press, 2012
7. Ngoc Lan, 2023. *AI making a mark in banking & finance*
8. Kris Sharma, 2023. *Machine learning in finance: history, technologies and outlook*
9. Lyn C Thomas, 2009. *Consumer Credit Models: Pricing, Profit and Portfolio*

1. Overfitting: Quá khớp có thể bị hình thành do 3 nhóm chính: Quá ít dữ liệu đầu vào, Mô hình quá phức tạp, Sự không cân bằng trong dữ liệu, [↑](#footnote-ref-1)
2. Model reliance đề cập đến mức độ mà một hệ thống hoặc ứng dụng phụ thuộc vào các mô hình Học máy để thực hiện các nhiệm vụ nhất định. Khi một hệ thống có mức độ reliance cao vào các mô hình Học máy, điều này có nghĩa là hiệu suất hoặc khả năng của hệ thống phụ thuộc nhiều vào độ chính xác và khả năng dự đoán của các mô hình đó. Nó cũng có thể ám chỉ đến việc mô hình chỉ hoạt động tốt trên một tập dữ liệu cụ thể mà nó được huấn luyện, và có thể không hiệu quả khi áp dụng vào các tình huống mới hoặc dữ liệu không gian khác. Để giảm bớt model reliance, các nhà nghiên cứu và các chuyên gia Học máy thường tập trung vào việc phát triển các mô hình có khả năng tổng quát hóa cao hơn, tức là mô hình có khả năng áp dụng được vào nhiều tình huống và tập dữ liệu khác nhau mà không mất đi hiệu suất quá nhiều. Điều này có thể đạt được thông qua việc sử dụng các kỹ thuật như regularization, data augmentation, và việc thu thập và sử dụng dữ liệu đa dạng hơn trong quá trình huấn luyện. [↑](#footnote-ref-2)
3. Sau khi trích lập dự phòng và write off, vẫn có một khối lượng tiền được khách hàng trả lại được ghi nhận trong recovery và không được phản ảnh trong mô hình này. [↑](#footnote-ref-3)