BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH

KHOA: CÔNG NGHỆ THÔNG TIN

LƯU THỊ YẾN NHI

KHOÁ LUẬN TỐT NGHIỆP

PHÂN TÍCH VÀ ỨNG DỤNG HỌC MÁY ĐỂ DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ THẺ TÍN DỤNG

Chuyên ngành: Khoa học dữ liệu

Giảng viên hướng dẫn: TS. Nguyễn Chí Kiên

TP. Hồ Chí Minh, tháng 05 năm 2023

INDUSTRIAL UNIVERSITY OF HO CHI MINH CITY

FACULTY OF INFORMATION TECHNOLOGY

LUU THI YEN NHI

GRADUATION THESIS

CREDIT CARD CHURN ANALYSIS AND PREDICTION BY USING MACHINE LEARNING

Major: Data Science

Instructor: Ph.D. Nguyen Chi Kien

Ho Chi Minh City, May 2023

CREDIT CARD CHURN ANALYSIS AND PREDICTION BY USING MACHINE LEARNING

ABSTRACT

Credit card customer churn is a major problem for financial institutions, as it leads to revenue loss and decreased customer loyalty. In this project, we propose a machine learning approach to predict credit card customer churn. We used a dataset of customer demographics, credit card usage patterns, and transaction history to build and evaluate several machine learning models, including Logistic Regression, Support Vector Machines, Random Forest, and Gradient Boosting. We also performed feature engineering and selection to improve the models' performance.

Our results show that Random Forest and Gradient Boosting are the most effective models for predicting credit card customer churn, with an accuracy of 95% and 97%, respectively. We also found that the most important features for predicting churn are customer age, credit limit, and payment history. These findings can be used by financial institutions to improve customer retention strategies and reduce revenue loss due to churn.

Overall, our project demonstrates the effectiveness of machine learning in predicting credit card customer churn and provides insights into the factors that contribute to churn. Future work could focus on improving the model’s performance by incorporating additional data sources or using more advanced techniques such as deep learning.

LỜI CẢM ƠN

Quá trình thực hiện luận văn tốt nghiệp là giai đoạn quan trọng nhất trong quãng đời mỗi sinh viên. Luận văn tốt nghiệp là tiền đề nhằm trang bị cho chúng em những kỹ năng nghiên cứu, những kiến thức quý báu trước khi lập nghiệp.

Trước hết, em xin phép gửi lời cảm ơn chân thành đến TS Nguyễn Chí Kiên, người đã giảng dạy, hỗ trợ , đồng hành cùng Khoá 15 Khoa học dữ liệu nói chung và trực tiếp hướng dẫn khoá luận tốt nghiệp nói riêng cho bản thân em. Nhờ thầy mà em có cơ hội làm việc trực tiếp trong môi trường thực tế về kinh tế, tài chính và từ các vấn đề ở doanh nghiệp để em có thể đúc kết và thực hiện luận văn này.

Ngoài ra em xin cảm ơn thầy Nguyễn Hữu Tình, giảng viên chủ nhiệm Khoá 15 Ngành Khoa học dữ liệu. Một người thầy tận tâm, tận tuỵ, tận tình đồng hành và đặt những viên gạch đầu tiên để ngành KHDL được thành lập. Thầy đã truyền lửa, truyền động lực để em có thể đặt quyết định vào ngành KHDL và hỗ trợ rất nhiều trong quá trình học tập, các buổi ngoại khoá, seminar, quá trình thực tập, làm luận văn để bản thân em được đúc kết thêm nhiều kiến thức, kinh nghiệm và cũng là hành trang đến với tương lai.

Em cũng xin cảm ơn quý Thầy, Cô khoa Công nghệ Thông tin. Đặc biệt là các Thầy, Cô ngành Khoa học dữ liệu đã tận tình chỉ dạy và trạng bị cho em những kiến thức cần thiết trong suốt quãng thời gian ngồi trên giảng đường. Em xin được bày tỏ lòng biết ơn đối với ban lãnh đạo trường Đại học Công Nghiệp Thành phố Hồ Chí Minh đã tạo điều kiện tốt nhất về cơ sở vật chất hay về tinh thần trong suốt quãng thời gian học tập tại trường.

Cuối cùng em xin cảm ơn anh chị, các bạn đã hỗ trợ cách riêng hay chung để em có thể hoàn thành luận văn một cách tốt nhất. Cảm ơn mọi người đã đồng hành cùng với em thời gian qua.

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày…. tháng …. năm ....

CHỮ KÝ CỦA GIẢNG VIÊN

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN PHẢN BIỆN 1

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày…. tháng …. năm ....

CHỮ KÝ CỦA GIẢNG VIÊN

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN PHẢN BIỆN 2

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày…. tháng …. năm ....

CHỮ KÝ CỦA GIẢNG VIÊN

MỤC LỤC

MỤC LỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

CHƯƠNG 1. GIỚI THIỆU

1.1. Lý do chọn đề tài

Hiện tại, thị trường năng động và cạnh tranh cao do có sẵn một số lượng lớn các nhà cung cấp dịch vụ, đặc biệt là các ngân hàng, trên toàn thế giới. Một trong những thách thức chính đối với lĩnh vực này là sự thay đổi trong hành vi của khách hàng. Khách hàng là cốt lõi của tất cả các ngành, đặc biệt là các tổ chức phụ thuộc vào khách hàng, chẳng hạn như ngành ngân hàng, chịu trách nhiệm nhận tiền gửi, đầu tư và cho vay. Khách hàng dài hạn được kết nối trực tiếp với việc tạo ra lợi nhuận; do đó, các ngân hàng nên tránh để mất khách hàng.

Harvard Business Review nghiên cứu việc có được một khách hàng mới đắt hơn từ 5 đến 25 lần so với việc giữ chân một khách hàng hiện có . Và tin rằng việc tăng tỷ lệ giữ chân khách hàng lên 5% có thể dẫn đến tăng lợi nhuận cho các công ty từ 25% đến 95% [1]. Do đó, cho rằng khách hàng là tài sản quan trọng nhất có tác động mạnh mẽ đến lợi nhuận của ngân hàng, có năm trụ cột thiết yếu cho hoạt động kinh doanh ngân hàng hiện đại: vốn, thanh khoản, rủi ro, tài sản và quản lý khách hàng. Tập trung hiệu quả vào năm trụ cột có thể đảm bảo rằng ban quản lý sẽ tối đa hóa lợi nhuận của ngân hàng một cách hiệu quả [2].

Do đó, sự rời bỏ của khách hàng là một thách thức cơ bản đối với các ngân hàng. Sự rời bỏ của khách hàng có thể được định nghĩa là việc mất khách hàng vào tay đối thủ cạnh tranh, dẫn đến tổn thất về lợi nhuận. Để quản lý việc rời bỏ, điều cần thiết là xác định những khách hàng có khả năng chuyển sang một ngân hàng cạnh tranh. Ngoài ra, Risselada et al. (2010) [3] đã chỉ ra rằng quản lý churning là rất quan trọng trong việc thiết lập mối quan hệ lâu dài phù hợp giữa các công ty và khách hàng để tối đa hóa giá trị của cơ sở khách hàng. Khách hàng rời bỏ có thể được chia thành hai nhóm: rời bỏ tự nguyện và không tự nguyện. Sự xáo trộn không tự nguyện xảy ra khi ngân hàng rút

dịch vụ từ khách hàng và rất dễ phát hiện. Mặt khác, sự rời bỏ tự nguyện khó xác định hơn, bởi vì đó là quyết định có ý thức của khách hàng khi chấm dứt mối quan hệ của họ với một ngân hàng nhất định. Hơn nữa, nó có thể được chia thành sự rời bỏ ngẫu nhiên và sự rời bỏ có chủ ý: sự rời bỏ ngẫu nhiên xảy ra khi có những thay đổi trong hoàn cảnh của khách hàng khiến họ không thể giao dịch với ngân hàng của mình (ví dụ: điều kiện tài chính) và điều này chiếm một tỷ lệ nhỏ [4][5]; sự rời bỏ có chủ ý là do nhiều yếu tố gây ra, bao gồm các dịch vụ công nghệ mới, giá cả tốt hơn và các yếu tố chất lượng.

Hầu hết các ngân hàng trên thế giới đều cung cấp dịch vụ thẻ tín dụng. Dịch vụ này khá phổ biến vì sử dụng thẻ tín dụng là một cách thuận tiện để thanh toán hóa đơn, hàng tạp hóa, tiền thuê nhà hoặc bất cứ thứ gì bạn cần. Thẻ tín dụng hoạt động tương tự như khoản vay ngắn hạn mà bạn phải thanh toán hóa đơn trước ngày đáo hạn khi kết thúc chu kỳ thanh toán.

Và hơn hết khách hàng luôn là vấn đề cốt lõi của mỗi doanh nghiệp, việc phân tích ***tỉ lệ rời bỏ (churn rate)*** đúng cách sẽ giúp cho công ty có được cái nhìn khá tổng quát về tình hình kinh doanh và hơn thế nữa. Từ thực tế khi tham gia thực tập doanh nghiệp cho thấy rằng ***tỉ lệ rời bỏ dịch vụ*** cung cấp tổng quát tình hình kinh doanh của doanh nghiệp cũng như những biến đổi bất thường (tốt hoặc xấu). ***Churn Rate*** giúp doanh nghiệp có cái nhìn rõ hơn về customer behavior, đi sâu vào phân tích vì sao khách hàng lại hủy hoặc ngừng sử dụng dịch vụ mình. ***Churn Rate*** giúp doanh nghiệp tìm ra được đâu là khách hàng quan trọng và những đối tượng nào chúng ta nên chú trọng vào; và cả cách tính Customer-lifetime-value (CLV).

Có thể thấy ***tỉ lệ rời bỏ/ ngừng sử dụng dịch vụ*** tuy không mới nhưng lại rất quan trọng trong thực tế phát triển doanh nghiệp, bất cứ công ty nào có kinh doanh trên hợp đồng hoặc thu phí hàng tháng đều có cần phân tích ***Churn Rate***. ***Tỉ lệ rời bỏ*** không chỉ

được dùng dựa trên số khách hàng *Churn* mà còn có thể dùng để phân tích lượng khách hàng "*Downgrade vs Upgrade*". Thậm chí cả những công ty về bán lẻ cũng có thể sử dụng nếu có các điều kiện nhất định kèm theo. Vì vậy tôi muốn đi sau vào phân tích, tìm hiểu về ***tỉ lệ rời bỏ*** và cụ thể hơn là bài toán phân tích tỉ lệ rời bỏ/ ngừng sử dụng thẻ tín dụng của khách hàng**.**

1.2. Mục tiêu nghiên cứu

Mục tiêu của bài toán này là tìm hiểu và phân tích về các yếu tố ảnh hưởng tới việc khách hàng không tiếp tục sử dụng thẻ tín dụng của ngân hàng.

Như đã nói ban đầu, Churn Rate có thể coi là một trong những chỉ số quan trọng hàng đầu trong các công ty subscriber-based. Đơn giản vì nguồn thu chính từ những công ty này đến từ phí sử dụng hàng tháng. Nếu lượng khách hàng hủy hợp đồng hay ngừng sử dụng dịch vụ hàng tháng cao hơn lượng khách hàng mới, đây chính là dấu hiệu của việc kinh doanh có vấn đề. Việc phân tích churn rate đúng cách sẽ giúp cho công ty có được cái nhìn khá tổng quát về tình hình kinh doanh và hơn thế nữa.

Ngoài ra mục tiêu của nghiên cứu này có thể đưa ra những kết quả dự đoán chính xác dựa trên việc huấn luyện dữ liệu đầu vào. Đưa ra các mô hình máy học phù hợp với các tính năng hàng đầu để dự đoán khách hàng rời bỏ và tìm cách tối ưu hoá các siêu tham số để mô hình đưa kết quả tốt nhất từ đó đưa ra kết luận và cái nhìn tổng quát về doanh nghiệp, tỉ lệ rời bỏ của khách hàng với sản phẩm thẻ tín dụng**.**

1.3. Đối tượng, phạm vi nghiên cứu

Churn Rate khái niệm không mới tuy nhiên ***Churn Rate*** có rất nhiều khía cạnh có thể nghiên cứu như: tỉ lệ gỡ ứng dụng, huỷ bỏ giao dịch, rời bỏ dịch vụ,vv…Ngoài ra đối với từng doanh nghiệp dữ liệu mỗi khác và đương nhiên dữ liệu cụ thể không được

công bố để có thể phân tích và nghiên cứu. Tôi gặp thách thức trong việc tìm bộ dữ liệu phù hợp để có thể nghiên cứu và may mắn khi có một bộ dữ liệu về Credit Card Customer Churn Prediction trong một cuộc thi cộng đồng được tổ chức vào khoảng năm 2019. Tập dữ liệu bao gồm hơn 10K khách hàng đề cập đến các thông tin sinh trắc học và hạn mức thẻ tín dụng, danh mục thẻ tín dụng tương ứng, v.v Có gần 18 tính năng và có khoảng hơn 16% khách hàng rời bỏ dịch vụ. Do đó hơi khó để đào tạo mô hình khi tính cân bằng dữ liệu chưa được đảm bảo.

Đối tượng nghiên cứu của bài toán là các khách hàng đã và đang sử dụng thẻ tín dụng. Mỗi dòng dữ liệu bao gồm các thông tin về nhân khẩu học, tình trạng sử dụng thẻ và các thuộc tính liên quan của mỗi khách hàng.

Phạm vi nghiên cứu của bài toán là tập trung vào tìm hiểu và phát triển các mô hình đánh giá khách hàng có thể ngừng sử dụng thẻ tín dụng hay không để đưa và các giải pháp hạn chế tối thiểu việc rời bỏ ấy, giảm thiệt hại cho ngân hàng.

1.4. Phương pháp nghiên cứu

Trước hết cần phân tích và khai phá dữ liệu để có thể hiểu rõ được các đặc trưng quan trọng của bộ dữ liệu từ đó có thể hiểu tình hình kinh tế và tài chính hiện tại, xác định các tác động của thị trường và xã hội đến khách hàng. Khai phá dữ liệu cho phép các doanh nghiệp có thể dự đoán được xu hướng tương lai. Quá trình khai phá dữ liệu là một quá trình phức tạp bao gồm kho dữ liệu chuyên sâu cũng như các công nghệ tính toán.

Tiếp theo sử dụng trí tuệ nhân tạo cụ thể là các mô hình học máy để phân tích và dự đoán hành vi khách hàng. Sử dụng trí tuệ nhân tạo có thể giúp phát hiện các mẫu và xu hướng khó nhận ra bằng mắt thường, từ đó đưa ra các dự đoán xu hướng khách hàng một cách chính xác và tổng quan hơn.

1.5. Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học

Giúp định lượng hóa và đánh giá các tác động tới khách hàng, từ đó đưa ra các quyết định và chiến lược để giữ chân khách hàng phù hợp. Nghiên cứu này cũng giúp phát triển các mô hình và công cụ để dự đoán hàng vi khách hàng rời bỏ thẻ tín dụng hiệu quả hơn. Nghiên cứu cũng giúp các tổ chức tài chính và ngân hàng hiểu rõ hơn về các vấn đề mà họ đang đối mặt và đưa ra các giải pháp phù hợp.

Ý nghĩa thực tiễn

Với việc nghiên cứu và phân tích ***Churn Rate***, doanh nghiệp có thể:

- Đánh giá "sức khỏe doanh nghiệp: và đưa ra các dự báo về tình hình kinh doanh trong tương lai.
- Xem xét các giá trị "thực" mà sản phẩm, dịch vụ mang lại cho người dùng và có sự điều chỉnh cần thiết từ sự biến động của ***Churn Rate***.
- Tính toán CLV (giá trị vòng đời khách hàng).
- Xác định tập khách hàng tiềm năng nhất công ty từ ***Churn Rate*** theo phân khúc khách hàng.

Ngoài ra, ***Churn Rate*** còn ảnh hưởng đến những chỉ số khác như:

- Monthly Recurring Revenue (doanh thu định kỳ hàng tháng): Khách hàng rời bỏ nhiều, tỷ lệ Churn cao đồng nghĩa với việc doanh nghiệp mất doanh thu, dẫn đến Monthly Recurring Revenue giảm.
- Customer Lifetime Value (giá trị vòng đời khách hàng): Tương tự như với Monthly Recurring Revenue, doanh thu trên mỗi vòng đời người dùng của công ty cũng sẽ giảm khi khách hàng rời bỏ.

1.6. Bố cục luận văn

Toàn bộ nội dung luận văn được trình bày trong 5 chương như sau:

Chương 1 – Tổng quan về lĩnh vực nghiên cứu Sơ lược tổng quan về vấn đề nghiên cứu trên phương diện tổng quan nhất, nêu ra mục tiêu, đối tượng nghiên cứu, phương pháp nghiên cứu và bố cục luận văn.

Chương 2 – Cơ sở lý thuyết và các nghiên cứu liên quan. Giới thiệu tổng quan về phân tích dữ liệu, về học máy cùng các vấn đề chung có thể gặp phải khi xây dựng mô hình dự đoán khả năng khách hàng rời bỏ thẻ tín dụng. Trình bày về một số nghiên cứu cùng với tình hình nghiên cứu trong nước và ngoài nước thời gian gần đây về đề tài đó.

Chương 3 – Mô hình đề xuất: Trình bày tổng quan về mô hình đề xuất và đi sâu phân tích các đặc trưng của mô hình đề xuất.

Chương 4 – Thực nghiệm trình bày các kết quả huấn luyện mô hình đạt được và phân tích, đánh giá.

Chương 5 – Kết luận và hướng phát triển.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Khai phá dữ liệu

Khai phá dữ liệu (Data Mining) là quá trình khám phá và phân tích các mẫu, thông tin tiềm ẩn và tri thức từ các dữ liệu lớn. Khai phá dữ liệu kết hợp các kỹ thuật và phương pháp trong các lĩnh vực như học máy, thống kê, trí tuệ nhân tạo, và cơ sở dữ liệu để tìm ra các mẫu, quy luật và thông tin tiềm ẩn trong dữ liệu.

Quá trình khai phá dữ liệu là một quá trình phức tạp bao gồm kho dữ liệu chuyên sâu cũng như các công nghệ tính toán. Hơn nữa, Data Mining không chỉ giới hạn trong việc trích xuất dữ liệu mà còn được sử dụng để chuyển đổi, làm sạch, tích hợp dữ liệu và phân tích mẫu.

Có nhiều tham số quan trọng khác nhau trong Data Mining, chẳng hạn như quy tắc kết hợp, phân loại, phân cụm và dự báo. Một số tính năng chính của Data Mining:

- Dự đoán các mẫu dựa trên xu hướng trong dữ liệu.
- Tính toán dự đoán kết quả
- Tạo thông tin phản hồi để phân tích
- Tập trung vào cơ sở dữ liệu lớn hơn.
- Phân cụm dữ liệu trực quan

2.2. Học máy

Học máy (Machine Learning) là một lĩnh vực khoa học máy tính đang phát triển nhanh chóng, tập trung vào việc phát triển các thuật toán và mô hình cho phép máy tính học hỏi từ dữ liệu và đưa ra dự đoán hoặc quyết định. Theo IBM [6], học máy là một nhánh của trí tuệ nhân tạo (AI) và khoa học máy tính tập trung vào việc sử dụng dữ liệu và thuật toán để bắt chước cách con người học, dần dần cải thiện độ chính xác của nó. Có thể chia hệ thống của các thuật toán học máy thành 3 thành phần chính:

- Quy trình quyết định (Decision process): Các thuật toán dựa trên dữ liệu đầu vào có thể gắn nhãn hoặc không gắn nhãn. Dựa trên dữ liệu đầu vào này thuật toán sẽ ước tính về mẫu dữ liệu từ đó đưa ra dự đoán hoặc phân loại.

- Hàm lỗi (Loss function): Loss function đánh giá dự đoán của mô hình. Nếu có các mẫu dữ liệu đã biết, loss function có thể so sánh để đánh giá độ chính xác của mô hình.
- Quy trình tối ưu hóa mô hình (Model optimization): Nếu mô hình có thể phù hợp hơn với các mẫu dữ liệu trong tập huấn luyện, thì các trọng số sẽ được điều chỉnh để giảm sự chênh lệch giữa dữ liệu đã biết và dữ liệu dự đoán của mô hình. Thuật toán sẽ lặp lại quy trình đánh giá, tối ưu hóa và cập nhật các trọng số của mô hình một cách tự động cho đến khi đạt đến ngưỡng tối ưu kỳ vọng.

Các thuật học máy được chia thành ba loại chính:

- Học máy có giám sát (Supervised machine learning): Supervised machine learning được xác định bằng cách sử dụng các bộ dữ liệu được gắn nhãn để huấn luyện các thuật toán nhằm phân loại dữ liệu hoặc dự đoán kết quả một cách chính xác.
- Học máy không giám sát (Unsupervised machine learning): Unsupervised machine learning sử dụng các thuật toán học máy để phân tích và phân cụm các bộ dữ liệu không được gắn nhãn. Các thuật toán này khám phá các mẫu hoặc nhóm dữ liệu ẩn mà không cần sự can thiệp của con người.
- Học máy bán giám sát (Semi-supervised machine learning): Trong quá trình đào tạo, semi-supervised machine learning sử dụng tập dữ liệu được gắn nhãn nhỏ hơn để huấn luyện khả năng phân loại phân loại của mô hình và trích xuất đặc trưng từ tập dữ liệu lớn hơn, không được gắn nhãn. Semi-supervised machine learning có thể giải quyết vấn đề không có đủ dữ liệu được gán nhãn cho thuật toán học có giám sát.

Hiện nay, đã có nhiều thuật toán học máy được nghiên cứu và phát triển ứng dụng. Một số thuật toán học máy thường được sử dụng như: Linear regression, Logistic regression, Clustering, Decision trees, Random forests, Neural networks.

Một trong những điểm mạnh chính của học máy là khả năng xác định các mẫu và mối quan hệ phức tạp trong các tập dữ liệu lớn mà con người khó hoặc không thể phát hiện được [7]. Học máy cũng có khả năng tự động hóa nhiều nhiệm vụ hiện đang được thực hiện bởi con người, giải phóng thời gian và nguồn lực cho các nhiệm vụ phức tạp hoặc sáng tạo hơn.

Tuy nhiên, cũng có những thách thức liên quan đến học máy, đặc biệt là liên quan đến các vấn đề về sai lệch, công bằng và minh bạch. Các thuật toán học máy chỉ tốt khi dữ liệu được đào tạo đầy đủ và đúng đắn. Nếu dữ liệu này chứa các sai lệch hoặc không chính xác, điều này có thể dẫn đến các dự đoán sai lệch hoặc không chính xác. Vì vậy, những nghiên cứu và ứng dụng cần nhận thực được những thách thức này và hướng tới phát triển các thuật toán mạnh mẽ, công bằng và minh bạch.

2.3. Bài toán dự đoán khách hàng ngừng sử dụng thẻ tín dụng

2.3.1. Tổng quan

Credit Card Churn là sự mất mát của một khách hàng sử dụng thẻ tín dụng, thông qua việc hủy bỏ hoặc không gia hạn tài khoản của họ. Chẳng hạn, khi thẻ tín dụng của bạn đến ngày hết hạn và bạn không gia hạn, bạn được coi là khách hàng bị rời bỏ. Tuy nhiên, sự gián đoạn cũng có thể xảy ra nếu bạn không sử dụng thẻ của mình trong một khoảng thời gian hoặc nếu bạn chuyển sang một công ty phát hành thẻ khác. Cuối cùng, nếu bạn chọn chủ động hủy tài khoản của mình, bạn cũng được coi là khách hàng rời bỏ.

Hiểu lý do tại sao khách hàng rơi vào bất kỳ danh mục nào trong số này là chìa khóa để dự đoán và ngăn chặn sự rời bỏ. Nếu khách hàng hủy tài khoản của họ, có thể là do dịch vụ khách hàng kém, thiếu phần thưởng, lãi suất cao hoặc một số yếu tố khác.

Mặt khác, nếu họ không sử dụng thẻ mới phát hành sau một khoảng thời gian nhất định, thì có thể là do họ chưa hiểu rõ về sản phẩm hoặc có thể họ chưa nhận được ưu đãi thích hợp để sử dụng thẻ. Các ngân hàng không hiểu được nguyên nhân khiến khách hàng rời bỏ có nguy cơ mất khách hàng và không thể thu hút khách hàng mới [8].

Theo đó, các ngân hàng nên thường xuyên theo dõi khách hàng để phát hiện các dấu hiệu cảnh báo về hành vi của khách hàng có thể dẫn đến sự rời bỏ. Hiện tại, các nhà nghiên cứu và quản lý ngân hàng nghiên cứu các mẫu và xu hướng trong dữ liệu để phát triển các mô hình có thể dự đoán liệu một khách hàng có kế hoạch rời bỏ hay không [9]. Hơn nữa, dữ liệu cung cấp các công cụ quan trọng trong ngân hàng; để khám phá các mẫu ẩn trong cơ sở dữ liệu lớn, nên áp dụng các quy trình phân cụm, bao gồm phân loại mạng thần kinh dựa trên các đặc điểm của khách hàng. Các quy trình này hỗ trợ xây dựng các mô hình dự đoán rời bỏ [10][11].

Dự đoán rời bỏ khách hàng sử dụng dữ liệu lớn là một lĩnh vực nghiên cứu trong công nghệ máy học, hoạt động để phân loại các loại khách hàng đặc biệt thành khách hàng rời bỏ hoặc không rời bỏ. Nhiều nghiên cứu trong tài liệu đã tạo ra nhiều mô hình dự đoán dựa trên các kỹ thuật thống kê và khai thác dữ liệu (mô hình học máy), chẳng hạn như hồi quy tuyến tính, cây quyết định, rừng ngẫu nhiên, hồi quy logistic, mạng lưới thần kinh, máy vectơ hỗ trợ và mạng lưới thần kinh sâu [12].

2.3.2. Các yếu tố ảnh hưởng đến chu kỳ thẻ tín dụng

Nhiều công ty thẻ tín dụng thấy khách hàng bỏ thẻ của họ với số lượng đáng báo động. Với mức phí cao, các chính sách trở nên phức tạp hơn và dịch vụ khách hàng thường không hữu ích, việc giữ thẻ trong ví của bạn trở nên khó khăn hơn. Một cuộc khảo sát mới của Bankrate đối với 2.582 người trưởng thành, bao gồm 2.301 chủ thẻ tín dụng, cho thấy 61% chủ thẻ người Mỹ đã hủy ít nhất một thẻ tín dụng. Ba mươi bảy phần

trăm trong số những chủ thẻ đó đã hủy nhiều hơn một thẻ. Đó là một thống kê đáng kinh ngạc khi xét đến sự tiện lợi và sức mua mạnh mẽ mà thẻ tín dụng mang lại [8].

Nhiều người trong số họ cho rằng chi phí cao là lý do số một khiến họ bỏ thẻ. Với các khoản phí trễ hạn, phí hàng năm, phí giao dịch nước ngoài và các chi phí ẩn khác, việc làm cho thẻ tín dụng xứng đáng với chi phí ngày càng trở nên khó khăn hơn.

Một số khách là cho rằng dịch vụ khách hàng là vấn đề lớn đối với thẻ tín dụng. Nhiều khách hàng bị mắc kẹt trong vòng lặp thời gian chờ đợi lâu hoặc thấy mình không thể nhận được câu trả lời rõ ràng cho câu hỏi của họ. Dịch vụ khách hàng kém có thể dẫn đến cảm giác bất lực và thất vọng, đồng thời có thể khiến khách hàng tìm kiếm trải nghiệm tốt hơn ở nơi khác.

Cuối cùng, hoàn cảnh cá nhân cũng đóng một vai trò. Trộm cắp danh tính là mối quan tâm lớn đối với nhiều người, khiến họ cảm thấy dễ bị tổn thương và bị lộ. Những người khác chỉ đơn giản là đã có đủ thẻ tín dụng và cảm thấy cần phải tách mình ra khỏi nợ nần [8][9].

2.3.3. Các nghiên cứu đã có

Dự đoán về sự rời bỏ của khách hàng thẻ tín dụng không phải là một lĩnh vực mới; nhiều nhà nghiên cứu đã phát triển các mô hình dự đoán khác nhau. Kaya và cộng sự. (2018) [13] đã phát triển một mô hình dự đoán xem xét hồ sơ giao dịch cá nhân của khách hàng. Mô hình chủ yếu sử dụng thông tin liên quan đến các yếu tố không gian, cũng như các yếu tố lựa chọn và đặc điểm hành vi. Kết quả cho thấy mô hình được phát triển có dự đoán chính xác hơn so với các mô hình truyền thống xem xét các đặc điểm dựa trên nhân khẩu học.

Hơn nữa, các nhà nghiên cứu ban đầu đã cố gắng giải quyết câu hỏi về cách phát triển các mô hình máy học để dự đoán sự rời bỏ của khách hàng, như đã thảo luận trong Miao và Wang (2022) [14]. Các tác giả đã phát triển một mô hình dự đoán sự rời bỏ của khách hàng sử dụng thẻ tín dụng bằng cách xem xét ba phương pháp học máy: rừng ngẫu nhiên, hồi quy tuyến tính và KNN. Tập dữ liệu được thu thập chứa 10.000 dữ liệu với 21 tính năng và mô hình được đánh giá bằng cách sử dụng ROC, AUC và Confusion matrix. Kết quả cho thấy Random Forest có hiệu suất tốt nhất so với các mô hình học máy khác, với độ chính xác 96,3%. Kết quả cho thấy ba biến quan trọng hàng đầu là tổng số tiền giao dịch, số lượng trong 12 tháng qua và tổng số dư quay vòng. Ngoài ra, de Lima Lemos et al. (2022) [15] đã điều tra những khách hàng bị rời bỏ trong lĩnh vực ngân hàng ở Brazil. Nghiên cứu nhằm tìm hiểu và dự đoán các biến số chính ảnh hưởng đến những khách hàng đã đóng hoặc dừng tài khoản của họ trong sáu tháng qua. Nghiên cứu đã sử dụng nhiều mô hình học máy khác nhau, kết quả cho thấy rừng ngẫu nhiên vượt trội so với các mô hình học máy khác ở một số chỉ số hiệu suất. Kết quả cho thấy những khách hàng có mối quan hệ bền chặt hơn với một tổ chức, những người vay ngân hàng nhiều hơn, ít có khả năng đóng tài khoản hơn.

Ngoài ra một nghiên cứu từ Dana AL-Najjar và công sự vào năm 2022 [16] đã phát triển một mô hình dự đoán sự rời bỏ của khách hàng sử dụng thẻ tín dụng bao gồm rừng ngẫu nhiên, mạng lưới thần kinh, Cây CR, cây C5, mạng Bayesian, cây CHAID, máy vectơ hỗ trợ, cây nhiệm vụ, hồi quy logistic đa thức và mô hình hồi quy tuyến tính. Để phát triển một mô hình dự đoán, ba phương pháp đã được sử dụng để kiểm soát số lượng biến độc lập được sử dụng trong mô hình dự đoán.

Model	Variables
Model 1	All variables
Model 2	All continuous variables and cluster value
Model 3	The selected variables after the feature-selection method

Hình 2.1: Mô hình đề xuất trong nghiên cứu của Dana AL-Najjar [17]

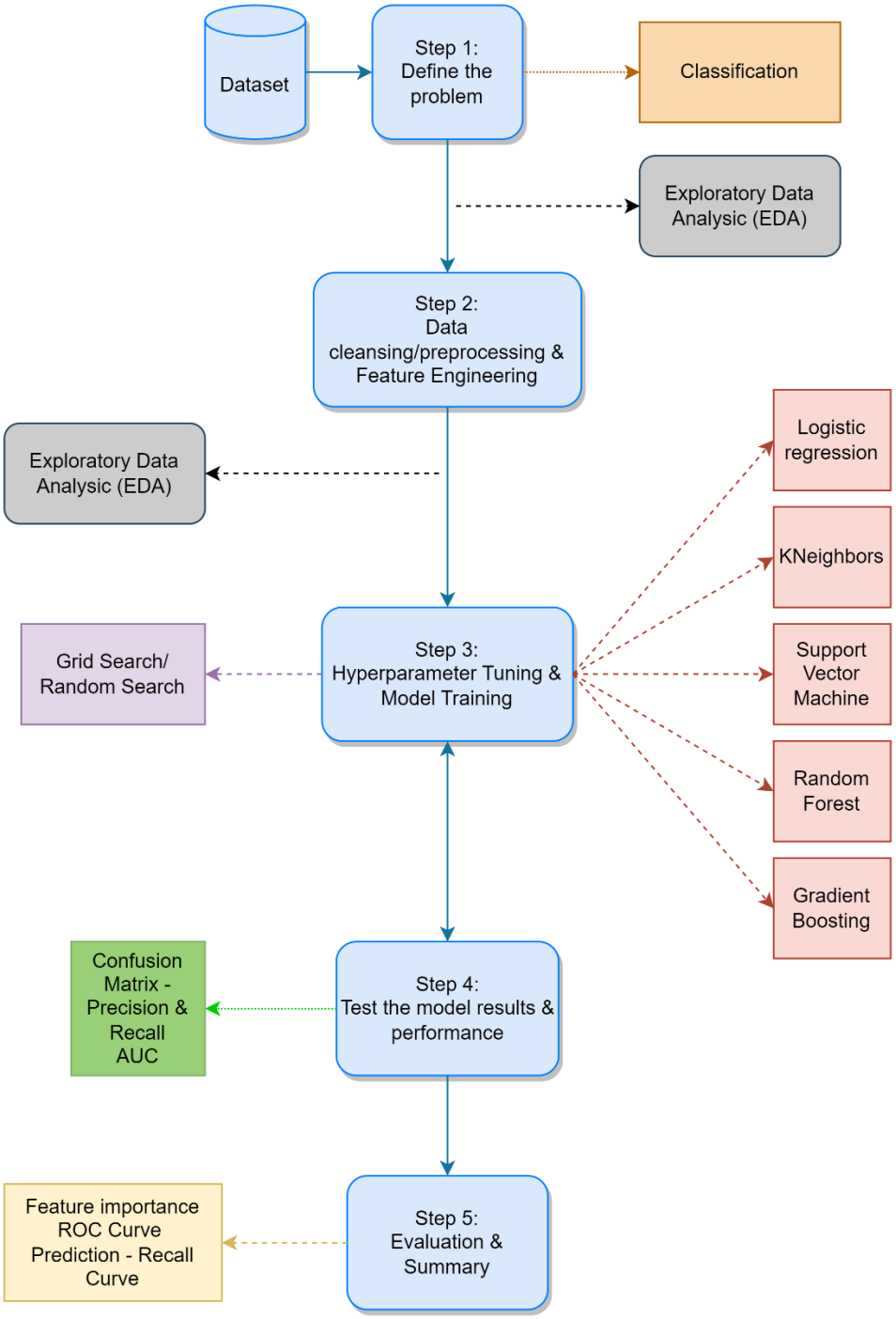
Đầu tiên, tất cả các biến độc lập được chuyển hướng đến một trong các mô hình máy học được áp dụng. Tiếp theo, họ đã cải thiện các biến độc lập bằng cách chỉ áp dụng phương pháp phân cụm hai bước cho các biến phân loại. Sau đó, các biến liên tục với các giá trị cụm được chuyển tiếp đến từng mô hình máy học. Để làm cho mô hình thực tế hơn, phương pháp hồi quy logistic đã được sử dụng để dự đoán số cụm dựa trên các biến phân loại. Và cho kết quả như hình dưới đây

	Models	Accuracy	Precision	Recall	FOR	F1_Score
Train	Bayesian Network	0.929	0.823	0.701	0.945	0.757
Train	C5 Tree	0.976	0.921	0.928	0.986	0.924
Train	CHAID	0.905	0.779	0.559	0.921	0.651
Train	CR-Tree	0.929	0.785	0.763	0.956	0.774
Train	Neural Network	0.925	0.830	0.667	0.939	0.739
Test	Bayesian Network	0.930	0.855	0.730	0.943	0.788
Test	C5 Tree	0.940	0.813	0.861	0.970	0.836
Test	CHAID	0.905	0.838	0.581	0.915	0.686
Test	CR-Tree	0.937	0.816	0.831	0.963	0.824
Test	Neural Network	0.926	0.839	0.723	0.942	0.777
Validate	Bayesian Network	0.926	0.821	0.662	0.941	0.733
Validate	C5 Tree	0.944	0.814	0.825	0.968	0.819
Validate	CHAID	0.902	0.752	0.544	0.921	0.631
Validate	CR-Tree	0.927	0.763	0.763	0.957	0.763
Validate	Neural Network	0.917	0.813	0.592	0.930	0.685

Hình 2.2: Kết quả dự đoán trong nghiên cứu của Dana AL-Najjar [17]

CHƯƠNG 3. MÔ HÌNH ĐỀ XUẤT

3.1. Tổng quan mô hình đề xuất



Hình 3.1: Mô hình đề xuất

Các bước để xây dựng mô hình cho bài toán dự đoán khách hàng rời bỏ thẻ tín dụng:

- Thu thập dữ liệu và định nghĩa bài toán.
- Tiền xử lý dữ liệu.
- Khai phá dữ liệu và khám phá đặc trưng.
- Xây dựng mô hình dự đoán với các tham số mặc định và đánh giá hiệu suất mô hình.
- Tối ưu hoá các siêu tham số.
- Đánh giá lại và kết luận.

3.2. Đặc trưng của mô hình đề xuất

3.2.1. Cân bằng dữ liệu

Trước khi chúng tôi đào tạo mô hình dự đoán của mình, trước tiên chúng tôi sẽ phải sửa lại tập dữ liệu mất cân bằng của mình. Chúng ta thấy rằng khách hàng rời bỏ chỉ chiếm 16,1% dữ liệu. Để giải quyết vấn đề này, chúng ta có thể sử dụng một số phương pháp, bao gồm:

- Undersampling: Phương pháp này sẽ loại bỏ một số tầng lớp đa số để tỷ lệ đa số và thiểu số bằng nhau hoặc ít nhất là không mất cân bằng (thường sử dụng tỷ lệ ngưỡng 2:1)
- Oversampling: Điều này sẽ nhân lên hoặc nhân đôi một số dữ liệu từ lớp thiểu số.
- Oversampling with SMOTE: Tương tự như quá trình lấy mẫu thông thường nhưng thay vì sử dụng cùng một điểm dữ liệu từ lớp thiểu số, SMOTE tạo dữ liệu tổng hợp mới từ lớp thiểu số.

SMOTE là viết tắt của Synthetic Minority Over-sampling Technique - một kỹ thuật để tăng cường mẫu dữ liệu cho lớp thiểu số trong bài toán phân loại. Đây là một trong những kỹ thuật được sử dụng để giải quyết vấn đề mất cân bằng dữ liệu trong bài toán phân loại, khi mà số lượng mẫu thuộc lớp thiểu số quá ít so với lớp đa số [18].

SMOTE hoạt động bằng cách tạo ra các mẫu dữ liệu nhân tạo cho lớp thiểu số. Cụ thể, kỹ thuật SMOTE sẽ chọn một mẫu dữ liệu từ lớp thiểu số và tìm kiếm k-nearest neighbors của nó. Sau đó, SMOTE sẽ tạo ra các mẫu dữ liệu mới bằng cách lấy các điểm nằm giữa các điểm trong k-nearest neighbors và mẫu dữ liệu ban đầu. Việc tạo ra các mẫu dữ liệu nhân tạo có thể được thực hiện bằng cách sử dụng các công thức toán học như sau:

- Chọn một mẫu dữ liệu từ lớp thiểu số.
- Chọn k mẫu dữ liệu gần nhất từ lớp thiểu số cho mẫu dữ liệu đã chọn.
- Với mỗi mẫu dữ liệu gần nhất, tạo ra một mẫu dữ liệu mới bằng cách lấy một tỉ lệ alpha ($0 < alpha < 1$) của khoảng cách giữa mẫu dữ liệu đã chọn và mẫu dữ liệu gần nhất đó.
- Tạo ra các mẫu dữ liệu mới bằng cách kết hợp các thuộc tính của mẫu dữ liệu đã chọn và các mẫu dữ liệu nhân tạo được tạo ra trong bước trên.
- SMOTE có thể được sử dụng để tăng cường mẫu dữ liệu cho lớp thiểu số và cải thiện hiệu suất của các mô hình phân loại. Tuy nhiên, nó cũng có thể dẫn đến overfitting nếu số lượng mẫu dữ liệu nhân tạo được tạo ra quá lớn. Do đó, việc lựa chọn k và alpha là rất quan trọng trong kỹ thuật SMOTE, và cần được điều chỉnh phù hợp để đạt được kết quả tốt nhất.

3.2.2. Logistic regression

Hồi quy logistic là một mô hình thống kê có thể được sử dụng để dự đoán xác suất xảy ra kết quả nhị phân (trong trường hợp này là liệu khách hàng sử dụng thẻ tín dụng có rời bỏ hay không). Trong dự đoán khách hàng rời bỏ thẻ tín dụng, hồi quy logistic có thể được sử dụng để xác định các yếu tố chính liên quan đến việc khách hàng rời bỏ, chẳng hạn như nhân khẩu học của khách hàng, mô hình sử dụng thẻ tín dụng và lịch sử giao dịch. Sau khi các yếu tố này được xác định, một mô hình logistic regression có thể được huấn luyện trên một tập dữ liệu các tương tác của khách hàng trong quá khứ để

dự đoán khả năng chuyển đổi cho khách hàng mới. Mô hình có thể được đánh giá dựa trên các chỉ số như accuracy, precision, recall, F1 score để xác định hiệu suất của nó. Nếu hiệu suất không đạt yêu cầu, mô hình có thể được cải thiện bằng cách điều chỉnh siêu tham số hoặc lựa chọn tính năng. Cuối cùng, mục tiêu là triển khai mô hình để đưa ra dự đoán chính xác về dữ liệu mới và ngăn chặn sự rời bỏ của khách hàng.

Đầu ra dự đoán của logistic regression:

$$f(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x}) \quad (3.1)$$

Trong đó:

- x là dữ liệu đầu vào.
- θ là hàm logistic.
- w là các tham số của thuật toán.

Đầu ra dự đoán của logistic regression là giá trị xác suất của biến mục tiêu thuộc về lớp positive (lớp dữ liệu quan trọng hơn cần được xác định đúng của bài toán),
với các thuộc tính dữ liệu đầu vào.
Công thức cho hàm mất mát của thuật toán hồi quy logistic:

$$\mathcal{J}(\theta) = -\left[\frac{1}{m} * \sum y^{(i)} \log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right] \quad (3.2)$$

Trong đó:

- $\mathcal{J}(\theta)$: Hàm mất mát.
- θ: Các tham số của mô hình.
- m: Số lượng mẫu dữ liệu.
- $y^{(i)}$: Giá trị đầu ra thực tế của mẫu dữ liệu $x^{(i)}$.
- $h_\theta\left(x^{(i)}\right)$: Giá trị đầu ra dự đoán của mẫu dữ liệu $x^{(i)}$.

Ưu điểm của mô hình:

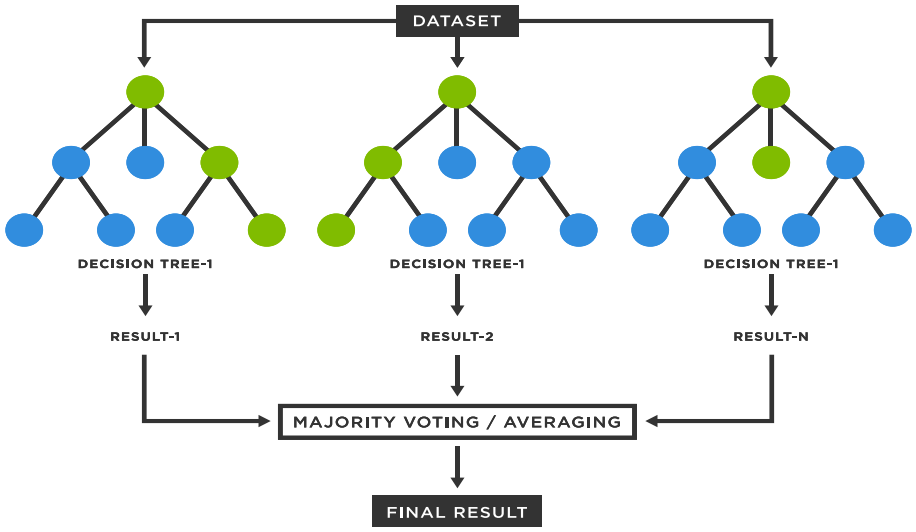
- Tính đơn giản và dễ hiểu: Hồi quy logistic là một thuật toán đơn giản và dễ hiểu. Nó cung cấp các hệ số rõ ràng cho từng tính năng đầu vào, cho phép dễ dàng giải thích tác động của từng tính năng đối với dự đoán rời bỏ. Điều này có thể có giá trị trong việc hiểu các yếu tố thúc đẩy sự rời bỏ của khách hàng.
- Giả định tuyến tính: Hồi quy logistic giả định mối quan hệ tuyến tính giữa các tính năng đầu vào và tỷ lệ cược log của việc rời bỏ. Mặc dù giả định này có thể không phải lúc nào cũng đúng trong các tình huống phức tạp, nhưng nó vẫn có thể mang lại kết quả hợp lý khi các mối quan hệ tương đối tuyến tính hoặc khi các tính năng được chuyển đổi phù hợp.
- Đầu ra xác suất: Hồi quy logistic cung cấp đầu ra xác suất, cho biết khả năng khách hàng rời bỏ. Xác suất này có thể hữu ích để xếp hạng khách hàng dựa trên khả năng rời bỏ của họ và ưu tiên các nỗ lực duy trì phù hợp.
- Hiệu quả: Hồi quy logistic hiệu quả về mặt tính toán và có thể xử lý các bộ dữ liệu lớn với nhiều tính năng. Nó ít tốn tài nguyên hơn so với một số thuật toán phức tạp hơn, làm cho nó phù hợp với các tình huống mà tài nguyên tính toán bị hạn chế.
- Xử lý dữ liệu không cân bằng: Dự đoán rời bỏ thẻ tín dụng thường liên quan đến các bộ dữ liệu mất cân bằng, trong đó số lượng khách hàng bị hủy bỏ nhỏ hơn nhiều so với khách hàng không bị đảo lộn. Hồi quy logistic có thể xử lý dữ liệu mất cân bằng bằng cách điều chỉnh trọng số của lớp hoặc sử dụng các kỹ thuật như lấy mẫu quá mức hoặc lấy mẫu dưới mức để cân bằng các lớp.
- Lựa chọn tính năng: Hồi quy logistic có thể giúp xác định các tính năng quan trọng nhất để dự đoán tỷ lệ rời bỏ bằng cách kiểm tra mức độ và ý nghĩa của các ước tính hệ số. Quá trình lựa chọn tính năng này có thể hỗ trợ giảm kích thước của tập dữ liệu và tập trung vào các yếu tố phù hợp nhất.
- Khả năng mở rộng: Logistic Regression chia tỷ lệ tốt cho các tập dữ liệu lớn và có thể xử lý cả vấn đề phân loại nhị phân và đa lớp. Nó có thể được áp dụng cho các dự án dự đoán sự rời bỏ thẻ tín dụng bất kể quy mô của cơ sở khách hàng.

3.2.3. Random Forest

Random Forest là một thuật toán học máy có thể được sử dụng cho các nhiệm vụ phân loại, chẳng hạn như dự đoán liệu một khách hàng sử dụng thẻ tín dụng có rời bỏ hay không. Nó hoạt động bằng cách xây dựng vô số cây quyết định tại thời điểm đào tạo và xuất ra lớp là chế độ của các lớp (phân loại) hoặc dự đoán trung bình (hồi quy) của từng cây. Cùng với thực hiện kỹ thuật tính năng và lựa chọn để xác định các tính năng quan trọng nhất để dự đoán sự rời bỏ của khách hàng thẻ tín dụng. Điều này có thể liên quan đến việc sử dụng các kỹ thuật như phân tích thành phần chính (PCA) hoặc phân tích tương quan để xác định các tính năng phù hợp nhất. Nhìn chung, Random Forest là một thuật toán học máy mạnh mẽ có thể được sử dụng để dự đoán khả năng rời bỏ thẻ tín dụng của khách hàng. Điều quan trọng là thử nghiệm các mô hình và phương pháp khác nhau để tìm ra mô hình và phương pháp tốt nhất [19].

Các bước tạo mô hình từ thuật toán Random forest:

1. Lấy **N** mẫu ngẫu nhiên với từ tập dữ liệu gốc để tạo mẫu bootstrap.
2. Huấn luyện mô hình decision trees trên mẫu bootstrap bằng cách sử dụng ngẫu nhiên **d** thuộc tính.
3. Lặp lại bước 1 và bước 2 để xây dựng **M** mô hình decision trees.
4. Khi đã xây dựng đủ số lượng mô hình decision trees đã đặt ra. Tiến hành dự đoán giá trị biến mục tiêu bằng cách tổng hợp các dự đoán của tất cả **M** mô hình decision trees với phương pháp lấy giá trị trung bình của các dự đoán (đối với mô hình dự báo) hoặc bằng cách sử dụng biểu quyết đa số (đối với mô hình phân loại).



Hình 3.2: Mô hình Random Forest

Nguồn: https://www.tibco.com/reference-center/what-is-a-random-forest

Ưu điểm của mô hình :

- Học tập theo nhóm: Random Forest là một thuật toán học tập theo nhóm kết hợp nhiều cây quyết định để đưa ra dự đoán. Cách tiếp cận tập hợp này giúp giảm tình trạng thừa và cải thiện hiệu suất tổng quát hóa của mô hình. Nó có thể xử lý các bộ dữ liệu nhiều chiều với các tương tác phức tạp giữa các tính năng, làm cho nó phù hợp để dự đoán thời hạn sử dụng thẻ tín dụng trong đó nhiều yếu tố có thể ảnh hưởng đến hành vi của khách hàng.
- Tầm quan trọng của tính năng: Random Forest cung cấp thước đo tầm quan trọng của tính năng, cho biết tính năng nào có tác động đáng kể nhất đến dự đoán rời bỏ. Thông tin này có thể giúp xác định các động lực chính khiến khách hàng rời bỏ và cung cấp thông tin chi tiết để đưa ra quyết định và các biện pháp can thiệp tiềm năng.
- Mối quan hệ phi tuyến tính: Random Forest có thể nắm bắt các mối quan hệ phi tuyến tính giữa các tính năng và biến mục tiêu, khuấy đảo. Nó có thể phát hiện

các mẫu và tương tác phức tạp mà các mô hình tuyến tính đơn giản hơn có thể không nắm bắt được. Trong dự đoán về sự rời bỏ thẻ tín dụng, thường có các mối quan hệ và tương tác phi tuyến tính giữa các thuộc tính, mô hình sử dụng và hành vi rời bỏ của khách hàng.

- Mạnh mẽ đối với các giá trị ngoại lệ và giá trị bị thiếu: Random Forest mạnh mẽ đối với các giá trị ngoại lệ và có thể xử lý các giá trị bị thiếu trong tập dữ liệu. Nó không dựa vào một cây quyết định duy nhất mà tổng hợp các dự đoán từ nhiều cây, giúp giảm thiểu tác động của dữ liệu bị nhiễu hoặc bị thiếu.
- Khả năng mở rộng: Random Forest có thể song song hóa, cho phép nó xử lý các tập dữ liệu lớn một cách hiệu quả. Nó có thể xử lý dữ liệu nhiều chiều với số lượng lớn các tính năng và không dễ bị ảnh hưởng bởi các biến không liên quan hoặc dư thừa.
- Khả năng diễn giải: Mặc dù các mô hình Random Forest không thể diễn giải được như các mô hình đơn giản hơn như hồi quy logistic, nhưng chúng vẫn cung cấp một số mức độ có thể diễn giải. Các thước đo tầm quan trọng của tính năng có thể giúp hiểu tính năng nào đóng góp nhiều nhất cho dự đoán rời bỏ, cung cấp thông tin chi tiết về hành vi của khách hàng.

3.2.4. Support Vector Machine (SVM)

Support Vector Machine (SVM) là một thuật toán học máy giám sát phổ biến được sử dụng cho phân loại, hồi quy và phát hiện ngoại lệ. SVM hoạt động bằng cách tạo ra một siêu phẳng trong không gian đa chiều sao cho tốt nhất phân chia các điểm dữ liệu vào các lớp khác nhau**. M**ục tiêu là tìm siêu phẳng có lề lớn nhất giữa hai lớp. Để tìm siêu phẳng này, chúng ta cần giải quyết một bài toán tối ưu hóa. Trong dự đoán rời bỏ khách hàng thẻ tín dụng, SVM có thể được sử dụng để tìm một siêu phẳng tách biệt tốt nhất hai loại khách hàng: những người sẽ rời bỏ và những người sẽ không. Siêu phẳng được chọn để tối đa hóa lề giữa hai lớp, điều này có thể cải thiện khả năng tổng quát

hóa của mô hình. SVM cũng có thể xử lý các mối quan hệ phi tuyến tính giữa các tính năng bằng cách sử dụng các hàm nhân để chuyển đổi dữ liệu thành không gian có chiều cao hơn.

Bài toán tối ưu hóa của SVM có thể được biểu diễn như sau:

$$\text{Minimize } \frac{1}{2}\|w\|^2 + C\sum(\xi_i) \tag{3.3}$$

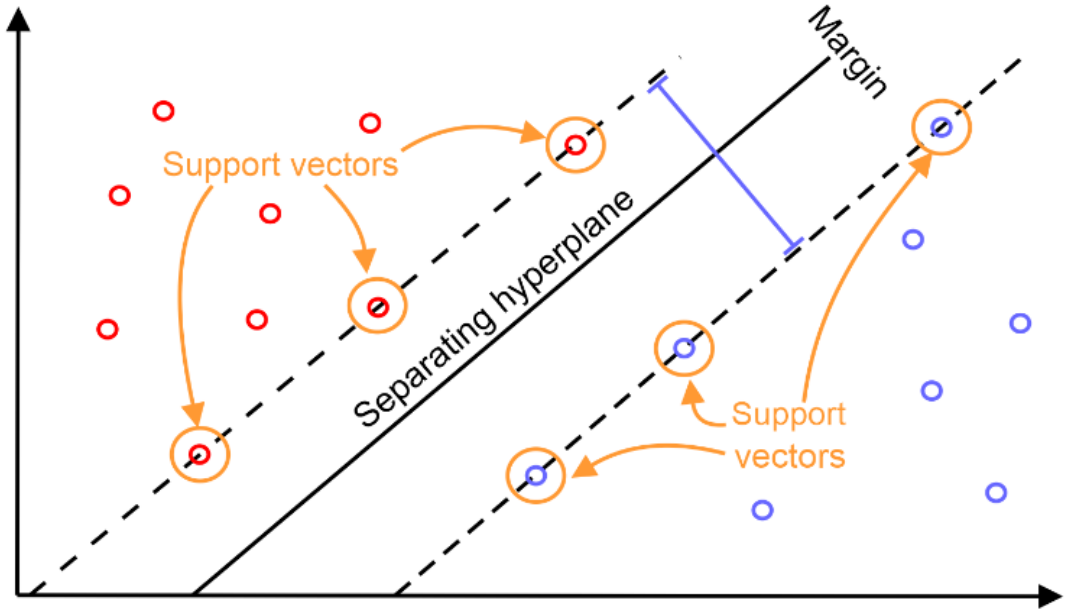
$$\text{Subject to } \quad \boldsymbol{y_i(wTx_i + b) \geq 1 - \xi_i, \forall_i = 1, \dots, n}$$

$$\boldsymbol{\xi_i \geq 0\ , \forall_i = 1, \dots, n}$$

Trong đó:

- $\boldsymbol{w}$ là vector trọng số của siêu phẳng.
- b là hệ số điều chỉnh độ lệch của siêu phẳng.
- $\mathbf{C}$ là tham số điều chỉnh độ quan trọng giữa việc tối đa hóa lề và việc tối thiểu hóa sai lầm phân loại. Nếu giá trị $\mathbf{C}$ lớn, SVM sẽ cố gắng tối đa hóa lề và chấp nhận một số điểm nằm trong vùng sai lệch (vi phạm ràng buộc). Nếu giá trị $\mathbf{C}$ nhỏ, SVM sẽ tối thiểu hóa số lượng điểm nằm trong vùng sai lệch và chấp nhận một lề nhỏ hơn.
- $\boldsymbol{y_i}$ là nhãn của điểm dữ liệu thứ i (yi=1 hoặc yi=-1).
- $\mathbf{x_i}$ là vector đặc trưng của điểm dữ liệu thứ i.
- $\boldsymbol{\xi_i}$ là biến lỏng giúp xác định điểm dữ liệu có nằm trong vùng sai lệch hay không.

Bài toán tối ưu hóa này có thể được giải bằng các phương pháp tối ưu hóa như hạ gradient (gradient descent), phương pháp Lagrange hoặc phương pháp lập trình toàn bộ (QP). Sau khi tìm được vector trọng số w và hệ số điều chỉnh độ lệch, ta có thể sử dụng chúng để phân loại các điểm dữ liệu mới. Cụ thể, một điểm dữ liệu mới sẽ được phân loại là thuộc lớp nào dựa trên khoảng cách của nó đến siêu phẳng. Nếu khoảng cách đến siêu phẳng lớn hơn lề, điểm dữ liệu được phân loại vào lớp thuộc siêu phẳng đó, trong trường hợp ngược lại, điểm dữ liệu được phân loại vào lớp khác [20].



Hình 3.3: Mô hình Support Vector Machine

Nguồn: svm-for-churn-prediction [20]

Ưu điểm của mô hình:

- Hiệu quả trong các không gian có nhiều chiều: SVM có thể hoạt động tốt trong các không gian có tính năng nhiều chiều, khiến nó phù hợp cho việc dự đoán thời hạn sử dụng thẻ tín dụng trong đó nhiều thuộc tính và hành vi của khách hàng được xem xét. Nó có thể xử lý các bộ dữ liệu với số lượng lớn các tính năng và vẫn duy trì hiệu suất tổng quát hóa tốt.
- Phân loại phi tuyến tính: SVM có khả năng mô hình hóa các mối quan hệ phi tuyến tính giữa các tính năng và churn. Bằng cách sử dụng các hàm hạt nhân như hạt nhân hàm cơ sở hướng tâm (RBF), SVM có thể nắm bắt một cách hiệu quả các mẫu phức tạp và ranh giới quyết định có thể tồn tại trong dữ liệu.
- Mạnh mẽ đối với việc trang bị quá mức: SVM có các cơ chế tích hợp sẵn để xử lý việc trang bị quá mức. Bằng cách điều chỉnh tham số chuẩn hóa (C), bạn có thể kiểm soát sự đánh đổi giữa tối đa hóa lề và giảm thiểu lỗi phân loại. Điều này

giúp ngăn chặn việc trang bị quá mức và cải thiện khả năng khái quát hóa của mô hình đối với dữ liệu chưa nhìn thấy.

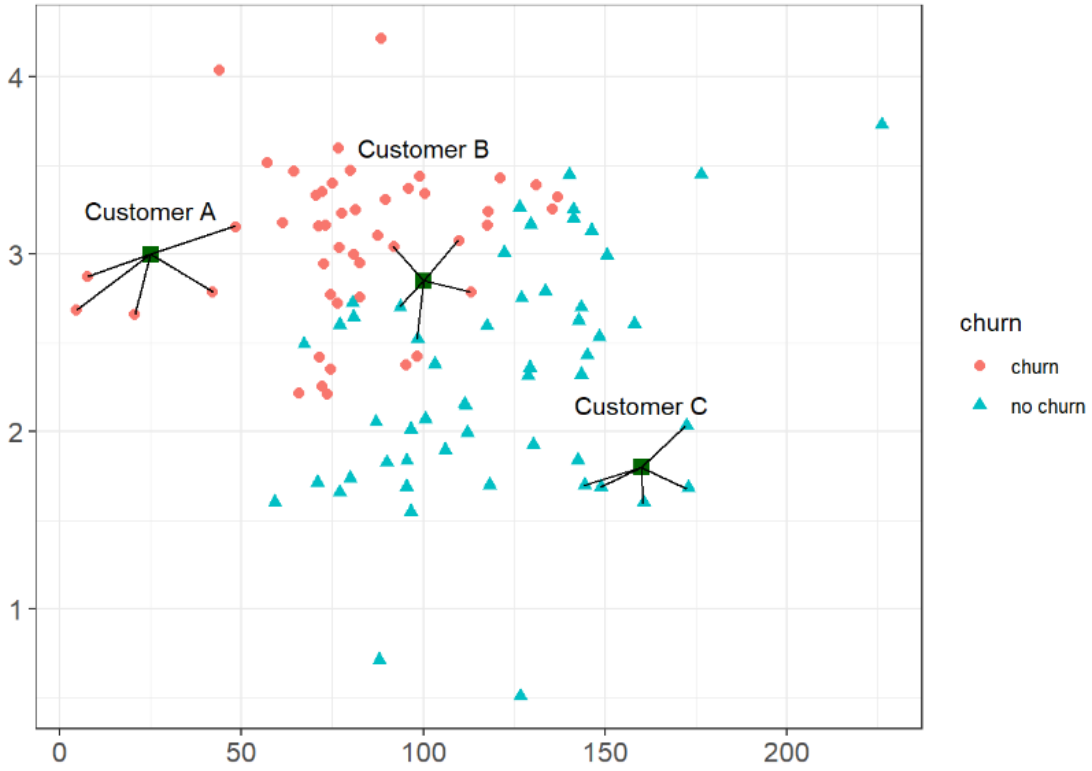
- Tối đa hóa biên: SVM nhằm mục đích tìm ranh giới quyết định tối đa hóa biên giữa các lớp. Thuộc tính tối đa hóa lề này cho phép SVM mạnh mẽ hơn đối với dữ liệu nhiễu và ngoại lệ, vì nó tập trung vào các mẫu gần ranh giới quyết định nhất.
- Xử lý dữ liệu không cân bằng: Trong dự đoán về sự rời bỏ thẻ tín dụng, người ta thường gặp phải các bộ dữ liệu mất cân bằng trong đó số lượng khách hàng bị rời bỏ nhỏ hơn nhiều so với những khách hàng không bị xáo trộn. SVM cung cấp các kỹ thuật như SVM theo trọng số lớp hoặc sử dụng học tập nhạy cảm với chi phí để xử lý dữ liệu không cân bằng và xem xét thích hợp cho lớp thiểu số.

3.2.5. K- Nearest Neighbors

K- Nearest Neighbors là một trong những thuật toán supervised-learning đơn giản nhất , thuật toán dựa trên khoảng cách, việc chia tỷ lệ đặc trưng là rất quan trọng để đảm bảo mỗi đặc trưng được đánh giá bằng cùng một phạm vi, giúp chúng có cùng trọng số trong tính toán khoảng cách giữa các điểm dữ liệu. Chọn số lượng hàng xóm gần nhất (K) để xem xét khi đưa ra dự đoán. Điều này có thể được thực hiện bằng cách thử nghiệm với các giá trị khác nhau của K và đánh giá hiệu suất của mô hình trên tập kiểm tra. Huấn luyện mô hình KNN trên tập huấn luyện bằng cách tính khoảng cách giữa mỗi điểm dữ liệu và K hàng xóm gần nhất của nó, và gán lớp phổ biến nhất của những hàng xóm đó cho điểm mới. Quá trình này được lặp lại cho mỗi điểm trong tập kiểm tra để đưa ra dự đoán.

Nhìn chung, KNN là một thuật toán đơn giản và hiệu quả để dự đoán sự rời bỏ thẻ tín dụng, đặc biệt là khi xử lý các tập dữ liệu vừa và nhỏ. Tuy nhiên, nó có thể không hoạt

động tốt trên các tập dữ liệu có nhiều tính năng hoặc các lớp không cân bằng cao, trong trường hợp đó có thể cần các thuật toán nâng cao hơn [21].



Hình 3.4: Mô hình K- Nearest Neighbors

Nguồn : https://bookdown.org/hbsabafaculty/ids_book/machine-learning-foundations.html

Trong không gian một chiều, việc đo khoảng cách giữa hai điểm đã rất quen thuộc: lấy trị tuyệt đối của hiệu giữa hai giá trị đó. Trong không gian hai chiều, tức mặt phẳng, chúng ta thường dùng khoảng cách Euclid để đo khoảng cách giữa hai điểm. Khoảng cách này chính là cái chúng ta thường nói bằng ngôn ngữ thông thường là đường chim bay. Đôi khi, để đi từ một điểm này tới một điểm kia, con người chúng ta không thể đi bằng đường chim bay được mà còn phụ thuộc vào việc đường đi nối giữa hai điểm có dạng như thế nào nữa.

Việc đo khoảng cách giữa hai điểm dữ liệu nhiều chiều, tức hai vector, là rất cần thiết trong Machine Learning. Chúng ta cần đánh giá xem điểm nào là điểm gần nhất của một điểm khác; chúng ta cũng cần đánh giá xem độ chính xác của việc ước lượng và trong rất nhiều ví dụ khác nữa. Và đó chính là lý do mà khái niệm norm (chuẩn) ra đời. Một số chuẩn thường dùng:

$$\left|\left|x\right|\right|_2 = \sqrt{x_1^2 + x_2^2 + \cdots x_n^2} \ với\ p = 2 \tag{3.4}$$

$$\left|\left|x\right|\right|_p = (|x_1|^p + |x_2|^p + \cdots |x_n|^p)^{\frac{1}{p}} \tag{3.5}$$

$$\left|\left|x\right|\right|_1 = |x_1| + |x_2| + \cdots |x_n| \tag{3.6}$$

$$\left|\left|A\right|\right|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} \tag{3.7}$$

Ưu điểm của mô hình:

- Tính phi tuyến tính: KNN là một thuật toán phi tham số, có nghĩa là nó không đưa ra bất kỳ giả định nào về phân phối dữ liệu cơ bản. Điều này làm cho nó phù hợp để nắm bắt các mối quan hệ phi tuyến tính giữa các tính năng và tỷ lệ rời bỏ. Trong dự đoán tỷ lệ rời bỏ thẻ tín dụng, mối quan hệ giữa các yếu tố khác nhau ảnh hưởng đến tỷ lệ rời bỏ thẻ tín dụng có thể không tuyến tính và KNN có thể nắm bắt các mẫu phức tạp như vậy.
- Tính linh hoạt: KNN là một thuật toán linh hoạt có thể xử lý cả vấn đề phân loại nhị phân và đa lớp, khiến thuật toán này có thể áp dụng cho các tình huống dự đoán rời bỏ khác nhau. Nó có thể xử lý các phân phối lớp không cân bằng và không yêu cầu các nỗ lực kỹ thuật tính năng hoặc tiền xử lý quan trọng.
- Ranh giới quyết định cục bộ: KNN xác định lớp của một điểm dữ liệu dựa trên các lớp của các lân cận gần nhất của nó. Cách tiếp cận này có thể có lợi khi các mẫu rời bỏ thể hiện sự phụ thuộc cục bộ, nơi các khách hàng ở gần nhau (trong

không gian tính năng) có xu hướng có hành vi rời bỏ tương tự. KNN có thể nắm bắt các ranh giới quyết định cục bộ này một cách hiệu quả.

- Giải thích trực quan: Quá trình ra quyết định của KNN dựa trên khái niệm về sự giống nhau hoặc khoảng cách giữa các điểm dữ liệu. Điều này làm cho nó tương đối dễ dàng để diễn giải và giải thích các dự đoán. Nó có thể cung cấp thông tin chuyên sâu về các tính năng hoặc đặc điểm góp phần khiến khách hàng được phân loại là khách hàng bị khuấy động hoặc không bị khuấy động.

3.2.6. Gradient Boosting

Gradient Boosting là một thuật toán máy học có thể được sử dụng để dự đoán tỷ lệ rời bỏ thẻ tín dụng của khách hàng. Thuật toán hoạt động bằng cách xây dựng vô số cây quyết định và kết hợp các dự đoán của chúng để đưa ra dự đoán cuối cùng, trong đó mỗi cây tiếp theo sẽ cố gắng sửa các lỗi của cây trước đó.

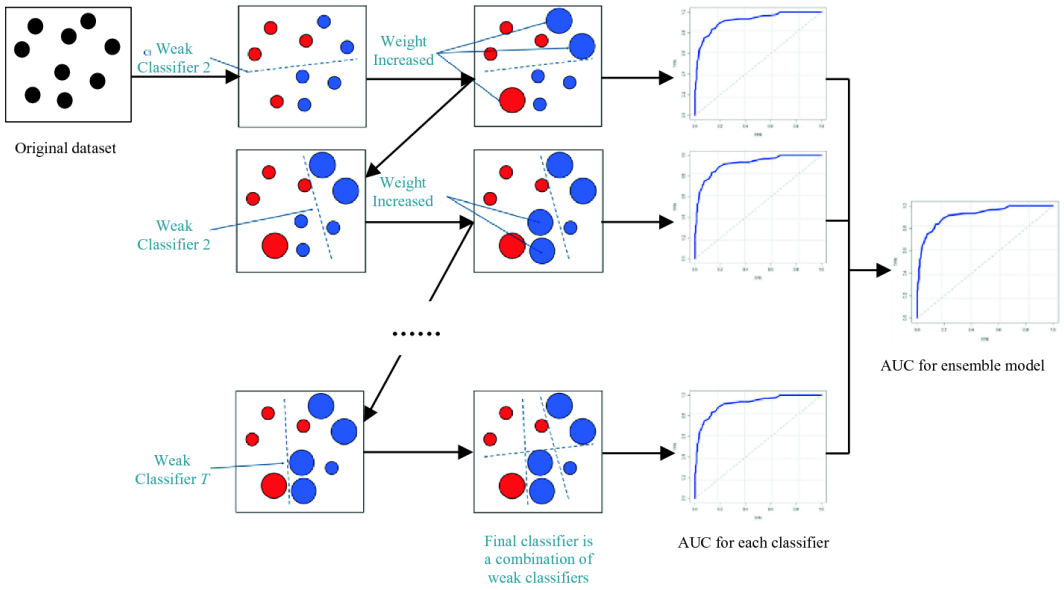
Với dữ liệu đầu vào $\mathbf{X}$ và biến mục tiêu là $\boldsymbol{y}$, thuật toán gradient boosting cố gắng tạo ra hàm $\widehat{\boldsymbol{f}}(\boldsymbol{x})$ với mục tiêu dự đoán. Tại mô hình thứ $\boldsymbol{b}$ trong chuỗi các mô hình dự đoán hàm mục tiêu tại mô hình đó là $\widehat{\boldsymbol{f}}^{\boldsymbol{b}}$. Mô hình tìm cách khớp giá trị phần dư $\mathbf{r^i}$ từ cây quyết định trước là $\widehat{\boldsymbol{f}}^{\boldsymbol{b-1}}$ **[22].**

Ưu điểm của mô hình:

- Khả năng dự đoán mạnh mẽ: Gradient Boosting được biết đến với độ chính xác dự đoán cao và khả năng xử lý các mẫu phức tạp trong dữ liệu. Nó tuần tự xây dựng một tập hợp những người học yếu (thường là cây quyết định) bằng cách tập trung vào các lỗi của những người học trước đó. Quá trình lặp đi lặp lại này cho phép Gradient Boosting học hỏi từ những sai lầm và cải thiện hiệu suất dự đoán tổng thể.
- Mối quan hệ phi tuyến tính: Gradient Boosting có thể nắm bắt hiệu quả các mối quan hệ phi tuyến tính giữa các thuộc tính của khách hàng và hành vi rời bỏ. Nó

có thể xác định các tương tác phức tạp và sự phụ thuộc giữa các tính năng, điều này rất quan trọng trong dự đoán sự rời bỏ thẻ tín dụng khi nhiều yếu tố có thể góp phần vào sự rời bỏ của khách hàng.

- Tầm quan trọng của tính năng: Gradient Boosting cung cấp thước đo tầm quan trọng của tính năng, cho biết tính năng nào có tác động đáng kể nhất đến dự đoán rời bỏ. Thông tin này có thể giúp xác định các động lực chính của việc rời bỏ và cung cấp thông tin chi tiết có giá trị cho việc ra quyết định và các chiến lược can thiệp tiềm năng.
- Xử lý dữ liệu không cân bằng: Dự đoán rời bỏ thẻ tín dụng thường liên quan đến các bộ dữ liệu mất cân bằng, trong đó số lượng khách hàng bị hủy bỏ nhỏ hơn nhiều so với khách hàng không bị đảo lộn. Thuật toán Gradient Boosting có thể xử lý dữ liệu mất cân bằng bằng cách sử dụng các kỹ thuật như hàm giảm trọng số hoặc phương pháp lấy mẫu để đưa ra sự cân nhắc thích hợp cho lớp thiểu số.
- Mạnh mẽ đối với các giá trị ngoại lệ và các giá trị bị thiếu: Các thuật toán Gradient Boosting thường mạnh mẽ đối với các giá trị ngoại lệ và có thể xử lý các giá trị bị thiếu trong tập dữ liệu. Họ có thể cung cấp dữ liệu bị thiếu bằng cách xem xét các phần tách thay thế hoặc sử dụng các chỉ số giá trị bị thiếu trong quá trình xây dựng cây.
- Học tập theo nhóm: Gradient Boosting là một kỹ thuật học tập theo nhóm kết hợp nhiều người học yếu để đưa ra dự đoán chính xác. Bằng cách tận dụng các điểm mạnh của nhiều mô hình, nó làm giảm việc trang bị thừa và cải thiện hiệu suất tổng quát hóa. Điều này làm cho nó rất phù hợp để dự đoán thời hạn sử dụng thẻ tín dụng, trong đó mục tiêu là khái quát hóa các mẫu và đưa ra dự đoán chính xác về dữ liệu không nhìn thấy được.



Hình 3.5: Mô hình Gradient Boosting

Nguồn : https://datascience.eu/machine-learning/gradient-boosting-what-you-need-to-know/

Gradient Boosting có base model là Decision Tree, ta biết đến 2 framework phổ biến nhất là XGBoost và LightGBM

3.2.6.1. XGBoost

XGBoost (Extreme Gradient Boosting) là một giải thuật được dựa trên gradient boosting, tuy nhiên kèm theo đó là những cải tiến to lớn về mặt tối ưu thuật toán, về sự kết hợp hoàn hảo giữa sức mạnh phần mềm và phần cứng, giúp đạt được những kết quả vượt trội cả về thời gian training cũng như bộ nhớ sử dụng[23].

XGBoost hoạt động bằng cách tạo một chuỗi các cây quyết định (decision trees), trong đó mỗi cây tiếp theo được tạo để sửa lỗi của các cây trước đó. Thuật toán có nhiều siêu tham số (hyperparameters) có thể được điều chỉnh để tối ưu hóa hiệu suất của thuật toán, bao gồm tốc độ học, số lượng cây và độ sâu tối đa của mỗi cây.

XGBoost đã trở thành một thuật toán phổ biến trong nhiều lĩnh vực khác nhau, bao gồm tài chính, chăm sóc sức khỏe, tiếp thị, …. Nó đã được sử dụng để giải quyết các vấn đề phức tạp, chẳng hạn như dự đoán các giao dịch gian lận, chẩn đoán tình trạng y tế và xác định khả năng rời bỏ của khách hàng. Tính linh hoạt, độ chính xác và khả năng mở rộng đã khiến XGBoost trở thành một trong những thuật toán học máy được sử dụng rộng rãi nhất hiện nay.

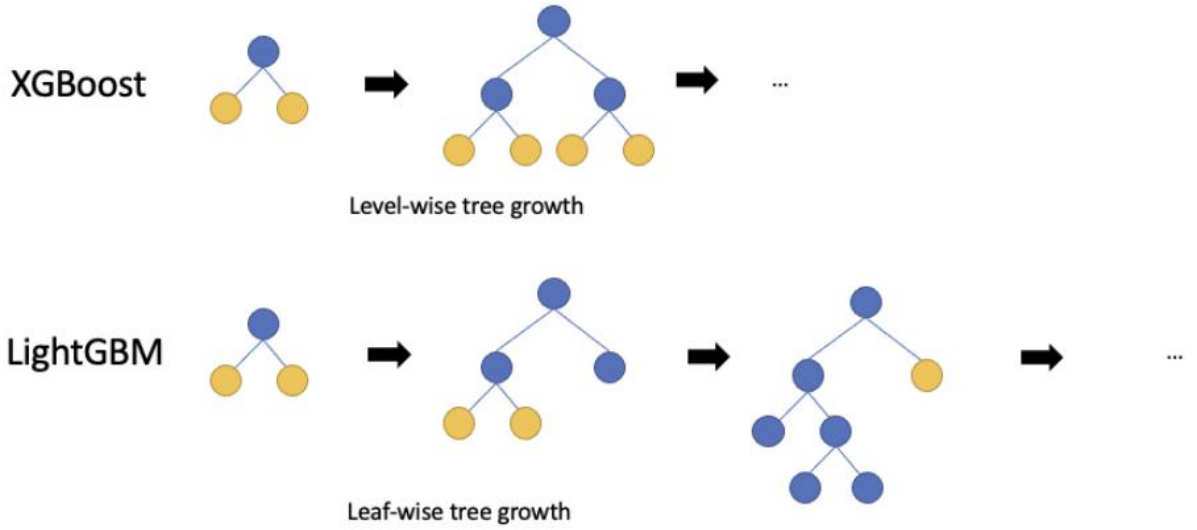
3.2.6.2. LightGBM

LightGBM là một khung gradient boosting mã nguồn mở sử dụng các thuật toán tree based learning. Nó được thiết kế để hoạt động hiệu quả, có thể mở rộng và chính xác, khiến nó trở thành lựa chọn phổ biến cho nhiều tác vụ học máy, bao gồm phân loại, hồi quy và xếp hạng. LightGBM được thiết kế để hoạt động hiệu quả với những ưu điểm sau [24][25]:

- Tốc độ đào tạo nhanh hơn và hiệu quả cao hơn.
- Sử dụng bộ nhớ thấp hơn.
- Độ chính xác tốt hơn.
- Hỗ trợ học song song, phân tán dựa trên quá trình sử dụng GPU.
- Có khả năng xử lý dữ liệu quy mô lớn.

Mô hình LightGBM dựa trên thuật toán cây quyết định tăng cường độ dốc (gradient boosting decision tree - GBDT), thuật toán này xây dựng một tập hợp các cây quyết định dự đoán biến mục tiêu bằng cách kết hợp các dự đoán của nhiều mô hình yếu hơn. Mỗi cây quyết định được xây dựng bằng cách lặp đi lặp lại quá trình thêm các nhánh mới vào cây hiện có với thuật toán tối ưu hàm mất mát. Thuật toán cũng kết hợp một kỹ thuật gọi là lấy mẫu một phía dựa trên độ dốc (Gradient-based one-side sampling - GOSS), GOSS lấy mẫu dữ liệu bằng cách ưu tiên các trường hợp có độ dốc lớn hơn, giúp giảm chi phí tính toán và cải thiện độ chính xác của mô hình [26][27].

Ngoài ra một điều giúp cho LightGBM có tốc độ đào tạo nhanh là do LightGBM thực hiện phân tách các nốt lá theo chiều dọc (leaf-wise) dẫn đến giảm thiểu tổn thất chi phí hơn các thuật toán phân chia nốt lá theo cấp độ (level-wise), chẳng hạn như thuật toán XGBoost thực hiện phân chia các nốt lá theo level-wise [28].



Hình 3.6: So sánh quá trình của thuật toán XGBoost và LightGBM

Nguồn : https://www.linkedin.com/pulse/xgboost-vs-lightgbm-ashik-kumar

Từ những yếu tố trên đã giúp LightGBM có tốc độ đào tạo nhanh nhưng vẫn giữ được chất lượng của mô hình đào tạo so với các thuật toán gradient boosting khác. Nhưng điều này cũng có thể làm cho mô hình gặp vấn đề overfitting tuy nhiên vấn đề này có thể được giải quyết bằng việc tối ưu độ sâu lớn nhất (max-depth) của mô hình [29].

3.2.7. Phương pháp đánh giá mô hình

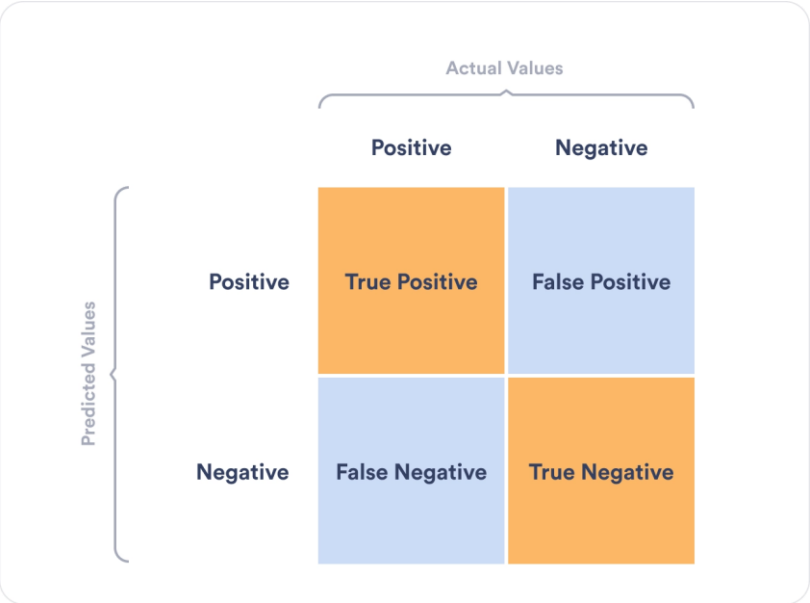
3.2.7.1. Confusion matrix

Trong học máy, các mô hình phân loại được sử dụng để phân chia dữ liệu thành các danh mục khác nhau. Theo Jianfeng và các cộng sự [30], Confusion matrix là phương pháp để tóm tắt hiệu suất của mô hình phân loại. Nó là một công cụ cần thiết để đánh giá độ chính xác của một mô hình và xác định mức độ phân loại dữ liệu, đặc biệt là các

mô hình đánh giá rủi ro tín dụng khi mà các dữ liệu thường có xu hướng mất cân bằng rất lớn [31][32] khiến cho kết quả của chỉ số đánh giá độ chính xác phân loại (classification accuracy) không còn đánh giá đúng chất lượng của mô hình huấn luyện [33].

Giả sử, với mô hình phân loại khách hàng rời bỏ thẻ tín dụng, nếu xem khách hàng rời bỏ là "positive" và khách hàng hiện tại được xem là "negative". Quá trình xây dựng confusion matrix cho mô hình phân loại 2 lớp bao gồm các bước sau:

1. Tạo tập dữ liệu kiểm thử với các biến mục tiêu phân loại ứng với các mẫu dữ liệu.
2. Đưa ra dự đoán cho từng mẫu dữ liệu trong tập dữ liệu kiểm thử bằng đầu ra của mô hình phân loại đã được huấn luyện.
3. Từ kết quả thật và kết quả dự đoán của các dữ liệu trong dữ liệu kiểm thử.
4. Đưa ra số lượng cho các lớp TP, TN, FP, FN. Trong đó:
    - TP (True Positive): Số lượng mẫu dữ liệu được mô hình dự đoán là "positive" và thực tế nhãn của mẫu dữ liệu là "positive".
    - TN (True Negative) : Số lượng mẫu dữ liệu được mô hình dự đoán là "negative" và thực tế nhãn của mẫu dữ liệu là "negative"
    - FP (False Positive): Số lượng mẫu dữ liệu được mô hình dự đoán là "positive" tuy nhiên nhãn thực tế của mẫu là "negative". (Những trường hợp dự đoán này được phân loại là sai lầm loại I).
    - FN (False Negative): Số lượng mẫu dữ liệu được mô hình dự đoán là "negative" tuy nhiên nhãn thực tế của mẫu là "positive" (Những trường hợp dự đoán này được phân loại là sai lầm loại II).



Hình 3.7: Confusion Matrix

Nguồn : https://plat.ai/blog/confusion-matrix-in-machine-learning/

Để biết mức độ chính xác của mô hình sẽ cần thêm một số chỉ số nhằm xác định hiệu suất phân loại của mô hình thông qua các chỉ số TP, TN, FP, FN. Các chỉ số đánh giá liên quan đến confusion matrix:

Accuracy: Tỷ lệ các dự đoán chính xác trên tổng số mẫu dự đoán.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (3.8)$$

Precision: Tỷ lệ của các kết quả dự đoán "positive" là đúng trên tổng số mẫu dự đoán là "positive". Trong mô hình phân loại sẽ kỳ vọng chỉ số này lớn nhất.

$$Precision = \frac{TP}{TP + FP} \quad (3.9)$$

Recall (Sensitivity - TPR): Tỷ lệ của các kết quả dự đoán là "positive" trên tổng số mẫu "positive" thực tế. Trong mô hình phân loại sẽ kỳ vọng chỉ số này lớn nhất.

$$Recall = \frac{TP}{TP + FN} \quad (3.10)$$

F1 score: Sử dụng "harmonic mean" để tính toán giữa 2 chỉ số recall và precision giúp việc so sánh giữa 2 mô hình có recall thấp, precision cao và ngược lại được thuận lợi hơn.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.11}$$

FPR: Tỷ lệ dự đoán sai các trường hợp "positive" trên tổng số mẫu dự đoán là "negative".

$$FPR = \frac{FP}{FP + TN} \tag{3.12}$$

Specificity: Tỷ lệ dự đoán đúng các trường hợp "negative" trên tổng số trường hợp "negative" thực tế

$$Specificity = \frac{TN}{FP + TN} = 1 - FPR \tag{3.13}$$

Confusion matrix cung cấp trực quan kết quả dự đoán của mô hình. Một trong những lợi ích chính của việc sử dụng confusion matrix là nó giúp xác định điểm mạnh và điểm yếu của mô hình. Bằng cách xem xét confusion matrix, có thể biết mô hình đang mắc phải loại lỗi nào và điều chỉnh cách tiếp cận phù hợp. Ví dụ: nếu chỉ số FP quá cao, thì cần điều chỉnh ngưỡng dự đoán để giảm số lượng dự đoán FP.

3.2.7.2. Cross-entropy

Cross-entropy là hàm loss được sử dụng mặc định cho bài toán phân lớp nhị phân. Nó được thiết kế để sử dụng với bài toán phân loại nhị phân trong đó các giá trị mục tiêu nhận một trong 2 giá trị {0, 1}. Về mặt toán học, nếu như MSE tính khoảng cách giữa 2 đại lượng số thì cross-entropy hiểu nôm na là phương pháp tính khoảng cách giữa 2 phân bố xác suất [34].

$$H(\boldsymbol{p}, \boldsymbol{q}) = -\sum_{i=1}^{C} p_i * log(q_i) \tag{3.14}$$

Trong đó C là số lượng các class cần phân lớp, trong bài toán binary classification thì C= 2.

3.2.7.3. AUC-ROC curve

AUC-ROC curve là một phương pháp tính toán hiệu suất của một mô hình phân loại theo các ngưỡng phân loại khác nhau. Với bài toán phân loại nhị phân kết quả đầu ra của mô hình sẽ là xác suất trong khoảng (0,1) và việc chọn ngưỡng để phân lớp đầu ra này rất quan trọng. Đường cong ROC (ROC curve) đồ thị của tỷ lệ TPR (3.10) so với tỷ lệ FPR (3.12) với các ngưỡng phân loại khác nhau. Để tạo ROC curve, xác suất dự đoán của mô hình cho lớp "positive" được sắp xếp theo thứ tự giảm dần. Sau đó, một ngưỡng dự đoán được đặt ra và tất cả các trường hợp có xác suất dự đoán lớn hơn hoặc bằng ngưỡng đó và được phân loại là "positive", trong khi các trường hợp khác được phân loại là "negative". TPR và FPR được tính toán dựa trên các phân loại này và quy trình được lặp lại cho các ngưỡng khác nhau [35] [36].

AUC là vùng bên dưới đường cong ROC, biểu thị xác suất mà một trường hợp "positive" được chọn ngẫu nhiên sẽ được xếp hạng cao hơn một trường hợp "negative" được chọn ngẫu nhiên theo mô hình. AUC nằm trong khoảng [0,1], với AUC = 0,5 cho biết mô hình không tốt và hoàn toàn không có khả năng phân loại giữa 2 lớp, AUC = 1 cho biết mô hình phân loại rất tốt [37].

CHƯƠNG 4. THỰC NGHIỆM VÀ KẾT QUẢ

4.1. Dữ liệu

4.1.1. Thu thập dữ liệu

Bài báo này phụ thuộc vào tập dữ liệu về sự rời bỏ của khách hàng thẻ tín dụng đối với các ngân hàng; tập dữ liệu được thu thập từ https://leaps.analytictica.com Khách hàng có tùy chọn để chọn một trong bốn loại thẻ tín dụng: blue, silver, gold or platinum. Khi khách hàng quyết định thay đổi ngân hàng của họ, họ được ghi nhận là khách hàng rời bỏ. Do đó, khách hàng rời đi khiến lợi nhuận của hệ thống ngân hàng giảm. Các chuyên gia ngân hàng ngày càng quan tâm đến việc thiết kế một hệ thống cảnh báo sớm để phân loại khách hàng của ngân hàng thành khách hàng rời bỏ hoặc không rời bỏ. Hệ thống sẽ có thể thông báo cho các nhà quản lý của ngân hàng để họ có thể liên lạc với những khách hàng dự kiến sẽ rời đi để cải thiện dịch vụ của họ, đây là một cách thích hợp để giữ cho khách hàng hài lòng với ngân hàng của họ. Bộ dữ liệu gồm 20 biến: 1 biến phụ thuộc và 19 biến độc lập. Tổng số khách hàng là 10.127, với 1627 khách hàng rời bỏ.

4.1.2. Tổng quan dữ liệu

Đầu tiên, bộ dữ liệu được chia thành các biến phân loại và biến liên tục dưới dạng các biến độc lập và một biến phụ thuộc (khách hàng rời bỏ). Tiếp theo, tôi đã phân tích các biến bằng cách sử dụng các số liệu thống kê khác nhau bao gồm min, max, phương sai, độ lệch chuẩn, chi bình phương (đối với biến phân loại) và phân tích tương quan (đối với biến liên tục)

Qua nhiều lần thực nghiệm và xử lý bộ dữ liệu bao gồm các bước: (1) Kiểm tra và xử lý các giá trị bị NaN; (2) Kiểm tra và xử lý các giá trị bất thường; (3) Mã hoá các biến phân loại; (4) Tìm hiểu mối tương quan giữa các dòng dữ liệu với cột Attrition_Flag, tổ chức lại bộ dữ liệu bằng các bước sau:

- Duplicates: Trước tiên hãy kiểm tra xem tập dữ liệu của có trùng lặp hay không và may mắn dữ liệu không có bất kỳ sự trùng lặp nào.
- Useless Features: Không phải tất cả các tính năng đều hữu ích trong phân tích, đó là lý do tại sao cần xóa các tính năng không cần thiết này này để đơn giản hóa bước tiếp theo. Có 3 tính năng thừa đó là:
  - `CLIENTNUM :` ID Khách hàng
  - `Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1` : Kết quả của một phân tích khác sử dụng Naive Bayes
  - `Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2 :` Giống như tính năng trước.

Sau khi tiền xử lý, tổ chức lại dữ liệu ban đầu còn các thuộc tính được định nghĩa như sau:

Tên cột	Ý nghĩa cột
Attrition_Flag	Trạng thái khách hàng (Existing Customer, Attrited Customer).
Customer_Age	Độ tuổi của khách hàng.
Gender	Giới tính (M=Male, F=Female)
Dependent_count	Số người phụ thuộc.
Education_Level	Trình độ học vấn.
Marital_Status	Tình trạng hôn nhân (Married, Single, Divorced, Unknown).
Income_Category	Thu nhập hàng năm (< 40K,40K - 60K, 60K–80K, 80K-120K, >120K, Unknown).
Card_Category	Loại thẻ (Blue, Silver, Gold, Platinum).

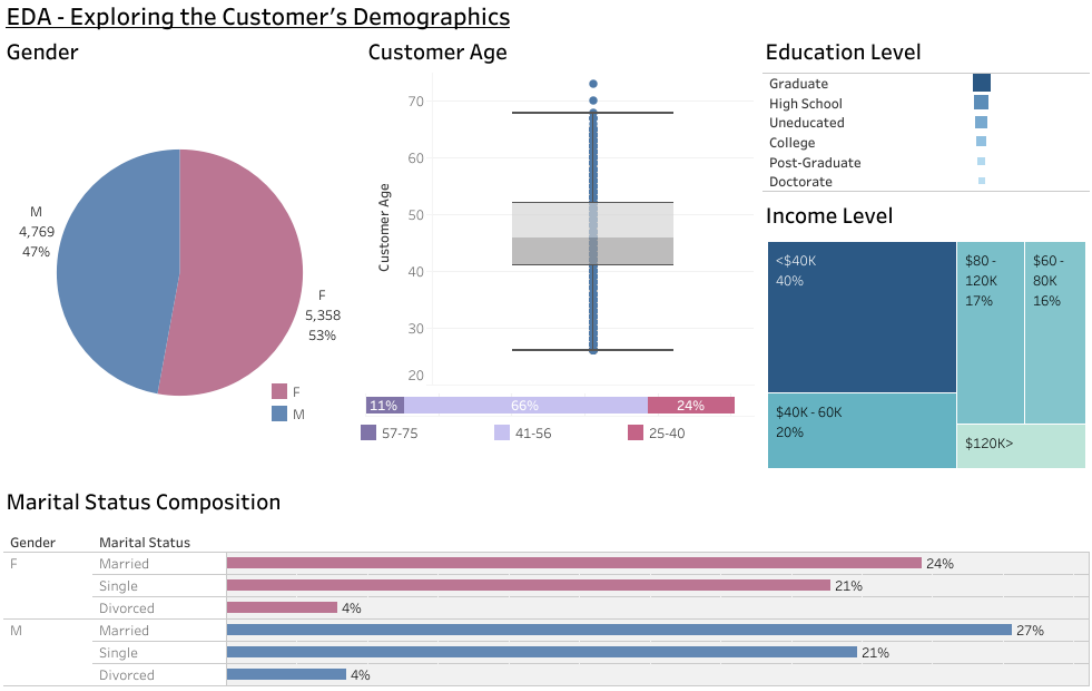
Months_on_book	Thời gian sử dụng dịch vụ của ngân hàng.
Total_Relationship_count	Tổng số sản phẩm khách hàng sử dụng.
Contacts_Count_12_mon	Số tháng không hoạt động trong 12 tháng qua.
Credit_Limit	Hạn mức tín dụng trên thẻ tín dụng.
Total_Revolving_Bal	Tổng số dư quay vòng trên thẻ tín dụng.
Avg_Open_To_Buy	Hạn mức tín dụng trung bình(Trung bình 12 tháng qua).
Total_Amt_Chng_Q4_Q1	Thay đổi về số tiền giao dịch (Q4 so với Q1).
Total_Trans_Amt	Tổng giá trị giao dịch (12 tháng qua).
Total_Trans_Ct	Số lượng giao dịch (12 tháng qua).
Total_Ct_Chng_Q4_Q1	Thay đổi về số lượng giao dịch (Q4 so với Q1).
Avg_Utilization_Ratio	Tỷ lệ sử dụng thẻ trung bình

Bảng 4.1: Tên và định nghĩa các thuộc tính của dữ liệu

4.1.3. Khai phá và phân tích dữ liệu (EDA)

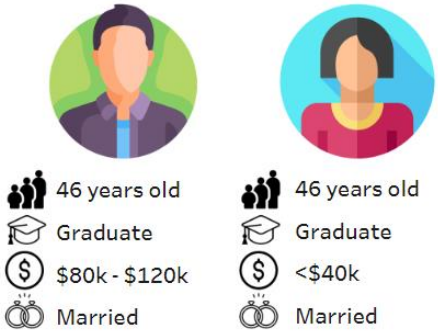
Trước khi chuyển sang phần học máy, tôi sẽ làm một chút về EDA trước. EDA là một khởi đầu tốt cho phân tích, nó giúp chúng tôi khám phá các mẫu, kiểm tra các giả định và phát hiện các điểm dữ liệu kỳ lạ như các giá trị ngoại lai. Tôi sẽ không hiển thị tất cả các biểu đồ và số liệu thống kê của từng tính năng, chỉ một số tính năng mà tôi cho là thú vị hoặc hữu ích cho các đề xuất kinh doanh.

4.1.3.1. Tổng quan dữ liệu



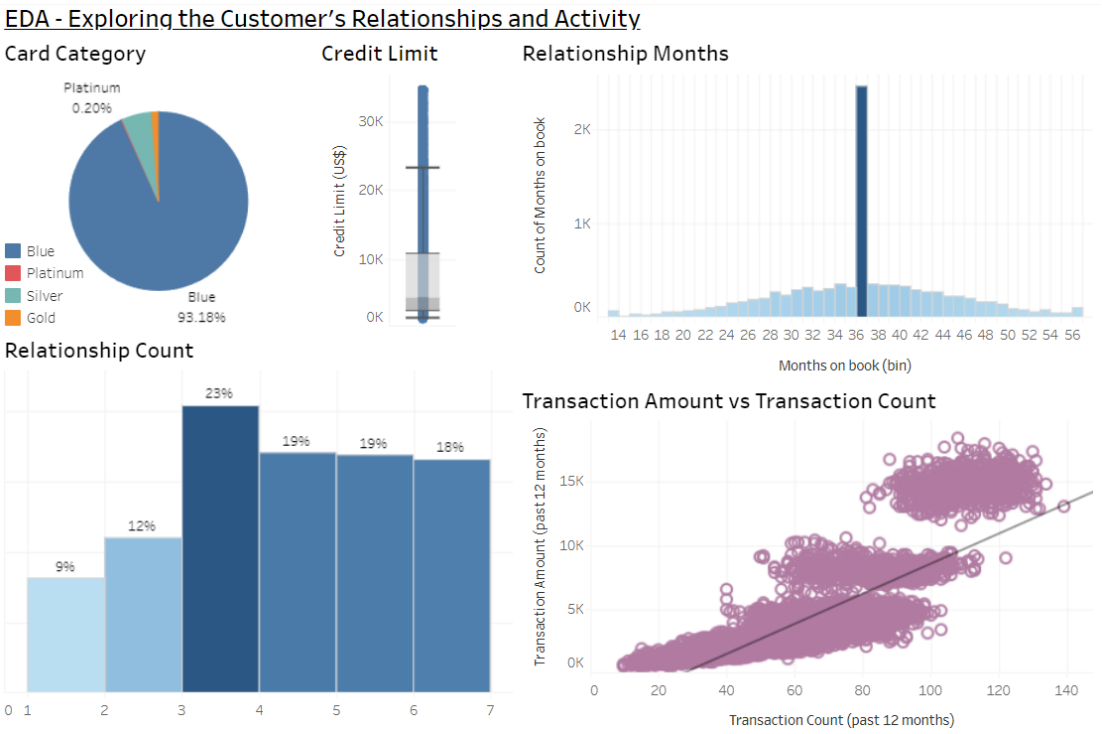
Hình 4.1: Tổng quan về nhân khẩu học của dữ liệu

- ***Income Category:*** Tỷ lệ khách hàng sử dụng thẻ tín dụng cao nhất đến từ những người có thu nhập dưới 40.000 USD. Đối tượng nhân khẩu học này là mục tiêu chính của tôi, vì vậy thay vì tập trung vào tiếp thị thẻ tín dụng cho những người có thu nhập cao, tôi có cơ hội thu hút khách hàng tốt hơn bằng cách quảng cáo cho những người có thu nhập dưới 40.000 USD.
- ***Education Level:*** Những người tốt nghiệp Đại học (Graduate) là thị trường lớn nhất của tôi. Cũng giống như thông tin chi tiết về danh mục thu nhập, chúng ta nên ưu tiên quảng cáo cho những người thuộc phân khúc này (Đã tốt nghiệp), điều này sẽ tăng tỷ lệ chuyển đổi của chúng ta.
- ***Age:*** Với 66% khách hàng trong độ tuổi từ 41-56 cần tập trung tiếp cận những ưu đãi để tiếp tục duy trì nhóm khách hàng tiềm năng này, và đương nhiên phải có chiến lược phù hợp với nhóm khách hàng Gen X này.



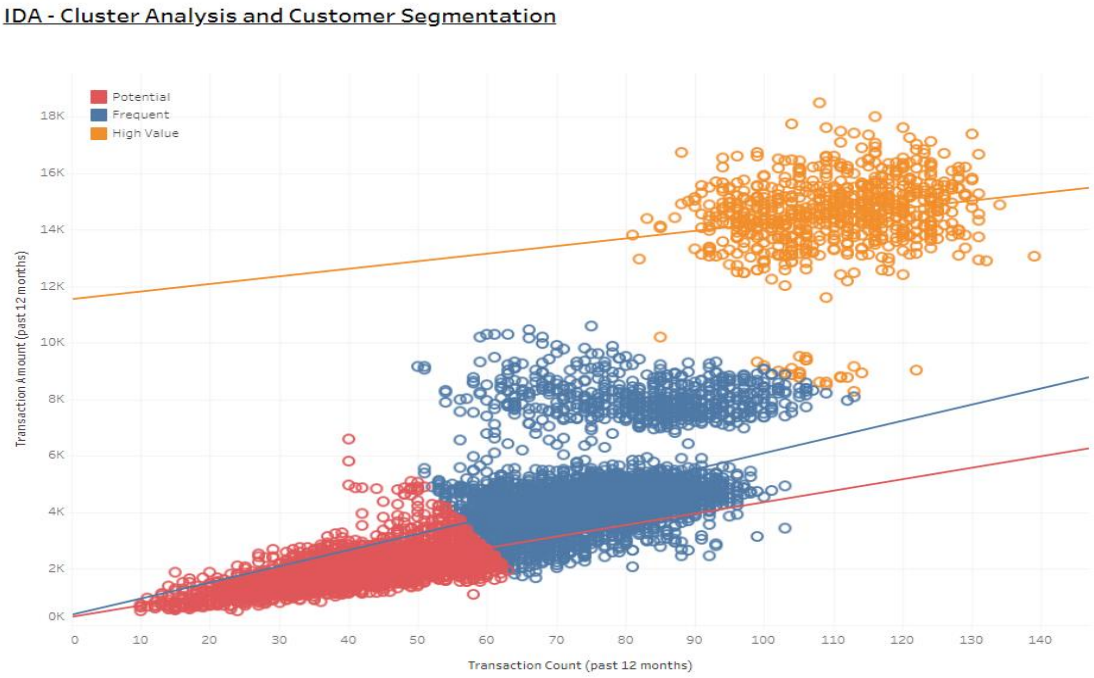
Hình 4.2: Khám phá các thuộc tính theo giới tính

- Ta có thể tóm tắt lại các thông tin về nhân khẩu học thu được như sau:
  - Đối với nam giới dữ liệu phổ biến có độ tuổi trung bình là 46 tuổi, đã tốt nghiệp, kết hôn và thu nhập trong khoảng từ $80k-$120k.
  - Đối với nữ giới dữ liệu phổ biến có độ tuổi trung bình là 46 tuổi, đã tốt nghiệp kết hôn và thu nhập nhỏ hơn $40k



Hình 4.3: Phân tích tỷ lệ khách hàng rời bỏ thẻ tín dụng

- ***Card Category:*** Đa số (93,18%) khách hàng đang giữ thẻ xanh cơ bản. Sự chênh lệch hiện ra rất rõ giữa các loại thẻ, cần có chiến dịch để thúc đẩy tập khách hàng đang ở mức thẻ cơ bản chiếm hơn 93% sử dụng những thẻ khác với mức ưu đãi và phân khúc phù hợp hơn.
- ***Credit Limit:*** Hạn mức tín dụng thường nằm trong khoảng từ $2.5k đến $10.8k, với mức trung bình là $4.3k.
- ***Relationship Months and Count*:** Hầu hết khách hàng đã gắn bó với ngân hàng được 36 tháng và có 3 thẻ của các ngân hàng khác nhau.
- ***Transaction Amount and Count:*** Quan sát thấy mối quan hệ tuyến tính, số lượng giao dịch tăng dẫn đến số tiền giao dịch tăng.

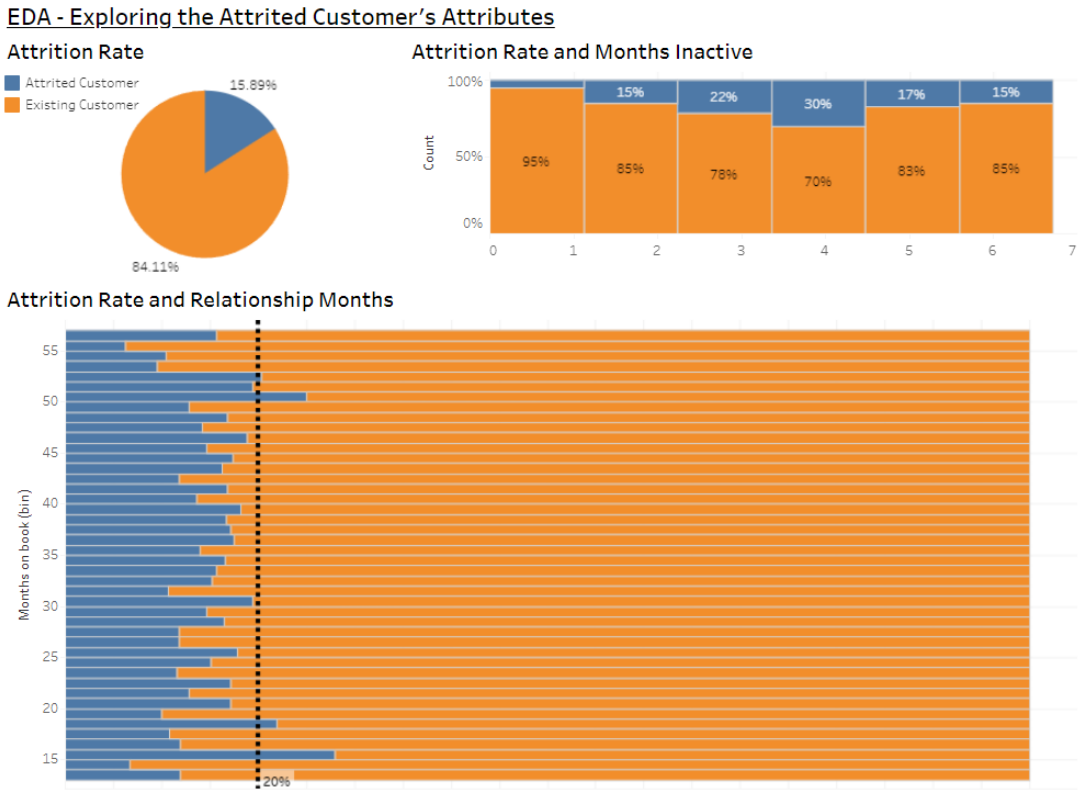


Hình 4.4: Phân tích cụm và phân khúc khách hàng

- Sử dụng phân tích cụm về giá trị và số lượng giao dịch có thể chia thành 3 phân khúc khách hàng cụ thể là (1) Tiềm năng, (2) Thường xuyên, (3) Khách hàng có giá trị cao đã được xác định.

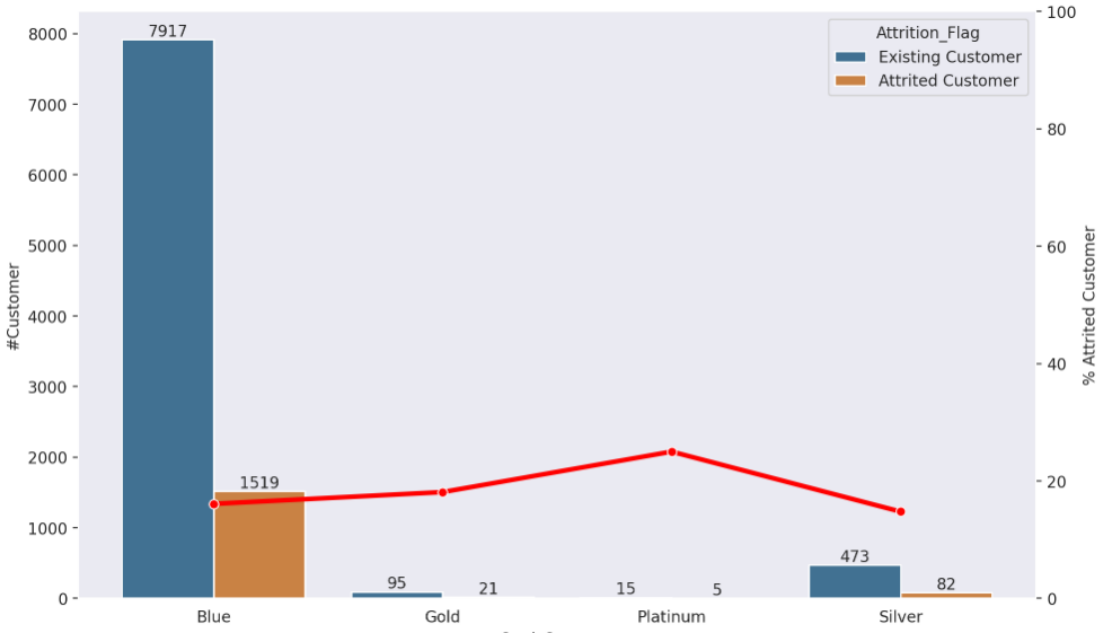
4.1.3.2. Chi tiết về dữ liệu khách hàng

- Nhìn chung, ở hình 4.5 ta thấy dữ liệu về khách hàng rời bỏ rất ít chỉ chiếm gần 16% trên tập dữ liệu.
- Hình 11 cũng có thấy việc ngừng sử dụng thẻ phổ biến khi khách hàng chưa sử dụng thẻ tín dụng trong 3 tháng và gắn bó cùng ngân hàng khoảng 15, 18, 50, 52 tháng (>= 20% tỷ lệ tiêu hao).
- Có thể thấy có nhiều khách hàng đã gắn bố với chúng ta rất lâu, điều gì đã khiến họ ngưng sử dụng thẻ? Ưu đãi không phù hợp? Hạn mức tín dụng không được thay đổi theo mức độ uy tín? Các ngân hàng khác có ưu đãi tốt hơn? Rất khó để có thể tìm kiếm tập khách hàng gắn bó với chúng ta lâu dài, nên thay vì đi tìm kiếm khách hàng mới cần đẩy mạnh để níu chân những khách hàng lâu năm để đưa đến khách hàng nhiều dịch vụ hơn của ngân hàng.



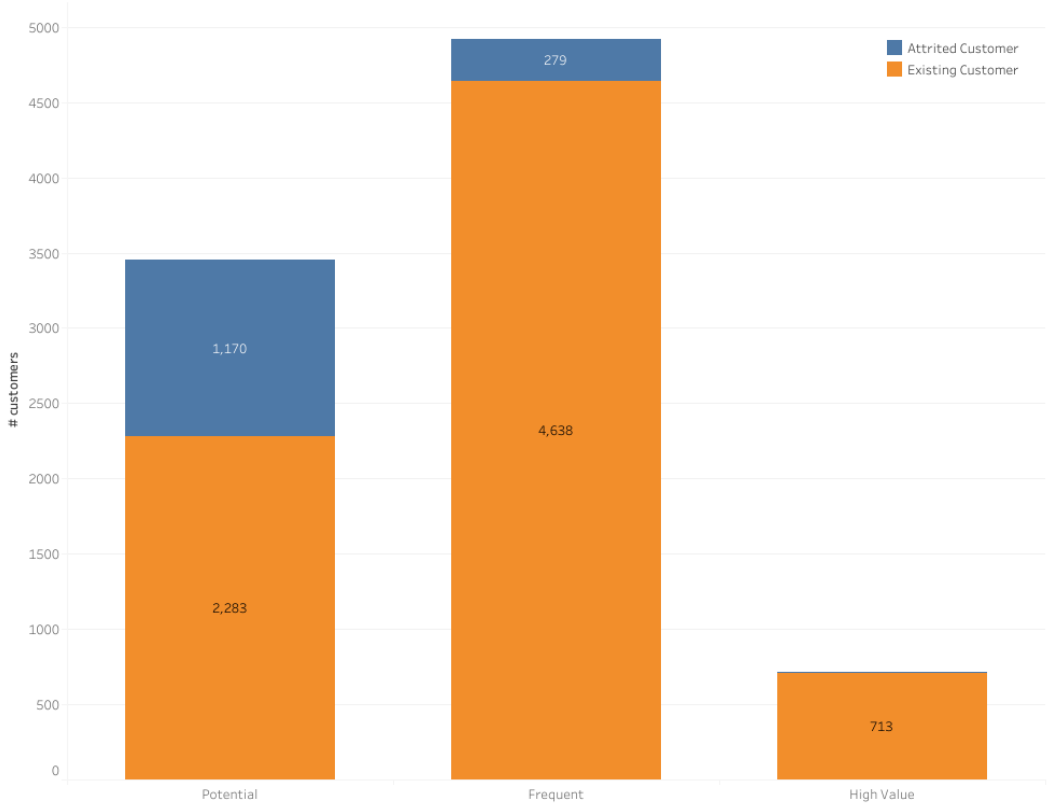
Hình 4.5: Thuộc tính của khách hàng đã và đang sử dụng

- Ta có thể thấy rằng thẻ Platinum có tới 25% tỷ lệ rời bỏ mặc dù số lượng khách hàng ở phân khúc này không nhiều nhưng điều này có thể có nghĩa là sản phẩm Platinum không đủ thỏa mãn khách hàng. Ta có thể tăng lợi ích của thẻ Platinum để giảm tỷ lệ tiêu hao.
- Ngoài ra số lượng rời bỏ của thẻ Blue rất lớn, hơn 1.5K chiếm 16% trên tổng khách hàng sử dụng thẻ Blue. Vì khách hàng tập trung chủ yếu ở loại thẻ này trong khi số lượng rời bỏ nhiều dẫn đến số lượng khách hàng sử dụng thẻ tín dụng giảm đáng kể, ảnh hưởng trực tiếp đến số lượng giao dịch, doanh thu, … Phần đa khách hàng hiện tại của ngân hàng đủ điều kiện mở thẻ Blue, việc triển khai nhiều khuyến mãi là điều cần thiết để khách hàng có thể sử dụng , trải nghiệm và gắn bó lâu hơn.



Hình 4.6: Phân phối loại thẻ của khách hàng

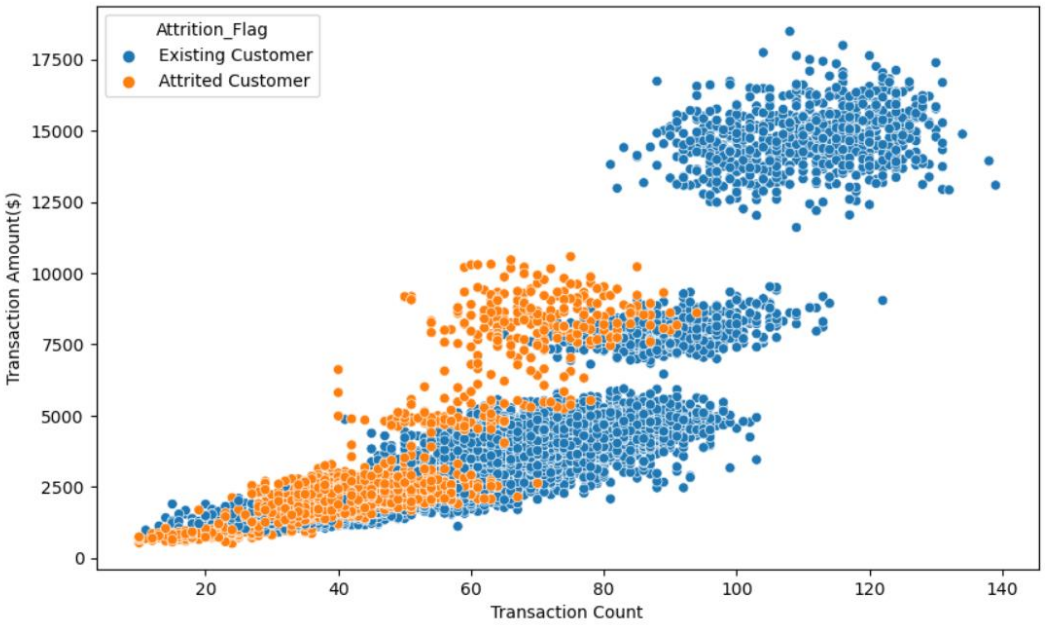
- Cụ thể ta có thế thấy 81% khách hàng được nhắm mục tiêu là phân khúc khách hàng tiềm năng trong khi 19% là phân khúc khách hàng thường xuyên. Điều may mắn là có rất ít sự tiêu hao đối với phân khúc khách hàng có giá trị cao.



Hình 4.7: Phân phối trên các cụm dựa vào tập khách hàng.

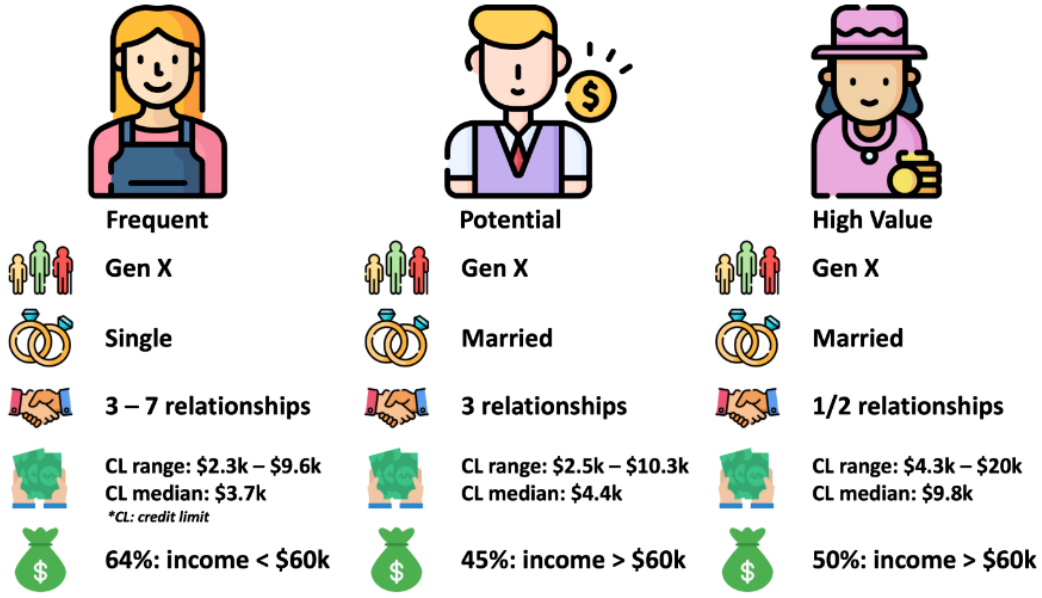
- Quan sát hình 4.8 chúng ta có thể thấy rằng có 3 cụm dựa trên số lượng giao dịch của họ so với số tiền giao dịch, điều này có thể được phân tích sâu hơn bằng cách sử dụng phương pháp học không giám sát nhưng chỉ từ biểu đồ này, chúng ta có thể thấy rằng những người có số lượng giao dịch và số lượng giao dịch cao có cơ hội tiêu hao thấp hơn , thậm chí không có 1 người nào được phân bổ/bị loại khỏi nhóm trên cùng bên phải. Các cụm giữa và cuối có một số người bị tiêu hao/bỏ đi, các cụm này có thể được phân tích và khảo sát thêm nếu có thể để giảm tỷ lệ tiêu hao của họ.

  Có thể thấy tập khách hàng có giá trị cao vẫn gắn bó sử dụng thẻ của ngân hàng, tuy nhiên vẫn có rất nhiều khách hàng thường xuyên sử dụng nhưng vẫn rời bỏ thẻ tín dụng. Cần phân tích và đưa ra giải phải để níu giữ tập khách hàng này vì họ đưa lại giá trị giao dịch cao và thường xuyên.



Hình 4.8: Phân cụm xu hướng khách hàng dựa trên giao dịch và giá trị giao dịch

4.1.3.3. Kết luận



Hình 4.9: Tổng quát và kết luận về dữ liệu

- Có thể tóm tắt tập dữ liệu như hình 15, chia thành 3 phân khúc khách hàng và cần có từng chiến dịch phù hợp đối với mỗi phân khúc tương ứng.
    - ***Frequent (thường xuyên)*** : Thuộc độ tuổi từ 41-56, tình trạng hôn nhân chủ yếu là độc thân, có từ 3-7 mối liên hệ với ngân hàng. Với hạn mức tín dụng từ $2,3K - $9,6K, trung bình khoảng $3,7K. Có tới 64% thu nhập dưới $60K
    - ***Potential (tiềm năng):*** Thuộc độ tuổi từ 41-56, tình trạng hôn nhân chủ yếu là đã kết hôn, có khoảng 3 mối liên hệ với ngân hàng. Với hạn mức tín dụng từ $2,5K - $10,3K, trung bình khoảng $4,4K. Có tới 45% thu nhập trên $60K
    - ***High Values (giá trị cao) :*** Thuộc độ tuổi từ 41-56, tình trạng hôn nhân chủ yếu là đã kết hôn, có khoảng 1/2 mối liên hệ với ngân hàng. Với hạn mức tín dụng từ $4,3K - $20K, trung bình khoảng $9,8K. Có tới 50% thu nhập trên $60K

4.1.4. Tiền xử lý dữ liệu

Sau khi loại bỏ một số tính năng không phù hợp bằng trực quan, ta đi sâu vào phân tích và xử lý bộ dữ liệu mới bằng các phương pháp.

4.1.4.1. Missing Values

Các giá trị bị thiếu sẽ làm giảm độ tin cậy của mô hình máy học, mô hình của chúng tôi có thể bị sai lệch. Đó là lý do tại sao chúng ta phải xử lý các giá trị còn thiếu trước khi tạo mô hình máy học. Có một số cách để xử lý các giá trị bị thiếu, bao gồm:

- Loại bỏ: Phương pháp này không được khuyến nghị đặc biệt nếu chúng ta không có nhiều điểm dữ liệu, nó có thể dẫn đến mất thông tin.
- Điền vào nó với median, mode hoặc mean: Phương pháp này khá đơn giản nhưng hãy sử dụng nó một cách thận trọng và đảm bảo rằng chúng ta chọn số liệu thống kê chính xác với lý do chính đáng.
- Điền nó với cùng một giá trị như các dữ liệu tương tự: Có thể tạo một hàm IF đơn giản theo cách thủ công tương quan với các tính năng khác hoặc bạn có thể tạo

một mô hình máy học để điền vào giá trị còn thiếu nhưng điều đó sẽ không thực tế và mất nhiều thời gian hơn.

- Điền vào nó với các giả định dựa trên kiến thức kinh doanh.
- Không có cách tốt nhất hoặc "chính xác" để xử lý các giá trị bị thiếu, tất cả phụ thuộc vào tập dữ liệu và tình huống, vì vậy hãy chọn phương pháp phù hợp với tập dữ liệu nhất.

4.1.4.2. Feature Encoding

Mã hóa tính năng được thực hiện để máy có thể đọc dữ liệu phân loại. Cho đến nay, máy móc chỉ có thể đọc các con số, đó là lý do tại sao chúng ta phải chuyển dữ liệu phân loại thành dữ liệu số. Có 2 loại mã hóa cơ bản:

- Ordinal Encoding: Tôi sử dụng mã hóa nhãn khi các đặc trưng có các giá trị thứ tự, chẳng hạn như trong tập dữ liệu của tôi, tôi có Card_Category, tính năng này có các giá trị thứ tự trong đó Blue Card là cấp thấp nhất và Platinum Card là cấp cao nhất.
- One-hot encoding: Sử dụng One-hot encoding cho tính năng khác không đáp ứng tiêu chí mã hóa nhãn. One-hot encoding sẽ tạo ra các tính năng/cột mới nhiều như số lượng giá trị duy nhất.

Ta sử dụng Mã hóa nhãn cho bất kỳ biến phân loại nào chỉ có 2 danh mục và Mã hóa một lần cho bất kỳ biến phân loại nào có nhiều hơn 2 danh mục.

4.1.4.3. Split Dataset

Để đảm bảo dữ liệu của chúng tôi không bị quá khớp, chúng tôi sẽ sử dụng phân tách thử nghiệm đào tạo đơn giản với tỷ lệ 80:20 tương ứng với (8101, 2026) khách hàng.

4.1.4.4. Cân bằng mẫu

Để cân bằng tập dữ liệu, chúng tôi sẽ sử dụng SMOTE. SMOTE (Kỹ thuật lấy mẫu tổng hợp thiểu số) bao gồm tổng hợp các yếu tố cho lớp thiểu số, dựa trên những yếu tố đã tồn tại. Nó hoạt động chọn ngẫu nhiên một điểm từ lớp thiểu số và tính k hàng xóm gần nhất cho điểm này. Các điểm tổng hợp được thêm vào giữa điểm đã chọn và các điểm lân cận của nó. Đầu tiên, chúng tôi chia dữ liệu đào tạo và thử nghiệm của mình.

SMOTE sẽ chỉ được áp dụng cho tập dữ liệu huấn luyện để dự đoán sẽ sử dụng tập dữ liệu gốc không có điểm dữ liệu tổng hợp.

Ta có thể thấy dữ liệu ban đầu mất cân bằng và với SMOTE điểm dữ liệu đã tăng lên cân bằng (cũng có thể lấy tỷ lệ ngưỡng 2:1).

Attrition_Flag	Customers	Customers SMOTE
Attrited Customer	1300	6801
Existing Customer	6801	6801

Bảng 4.2: Dữ liệu sau khi cân bằng mẫu bằng SMOTE

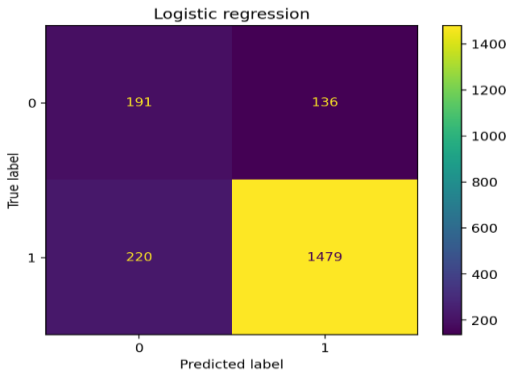
4.2. Thực nghiệm và đánh giá mô hình

4.2.1. Thực nghiệm với các tham số mặc định

Tôi đã thử một số thuật toán trong dự án này bao gồm Logistic Regression, KNeighbor Classifier, Random Forest, Support Vector Machine, XGBoost, and LightGBM. Trước hết tôi thử các thuật toán với các tham số ngẫu nhiên và đánh giá với bài toán đưa ra.

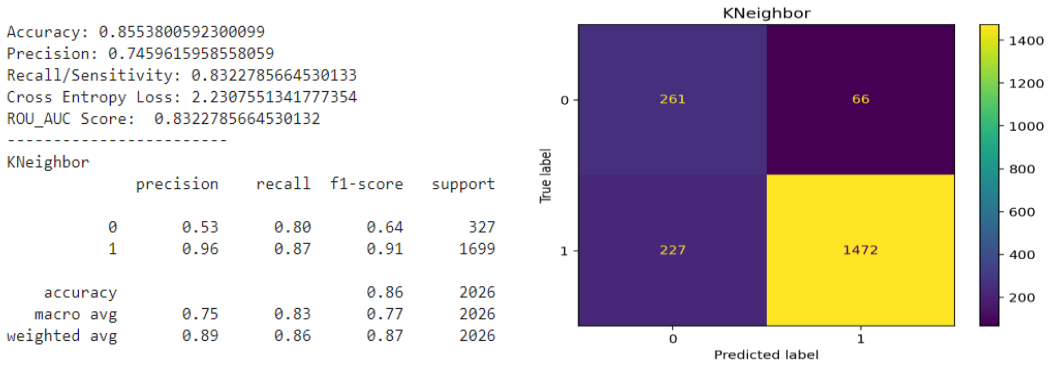
- Logistic Regression: Mô hình Hồi quy Logistic, có tới 191 dữ liệu với trạng thái Existing Customers đã được dự đoán chính xác và tối đa 1479 dữ liệu với trạng thái Attrited Customers đã được dự đoán chính xác, nhưng mô hình này cũng đưa ra dự đoán lỗi của 356 dữ liệu. Dựa trên những kết quả này, hiệu suất của mô hình thu được Accuracy Score là 82%, Precision score là 69%, Recall score là 73% với Cross Entropy Loss là 2,2079.

```
Accuracy: 0.824284304047384
Precision: 0.6902548341657062
Recall/Sensitivity: 0.7273049626241737
Cross Entropy Loss: 2.7104055555269553
ROU_AUC Score:  0.7273049626241735
------------------------
Logistic regression
              precision    recall  f1-score   support

           0       0.46      0.58      0.52       327
           1       0.92      0.87      0.89      1699

    accuracy                           0.82      2026
   macro avg       0.69      0.73      0.71      2026
weighted avg       0.84      0.82      0.83      2026
```



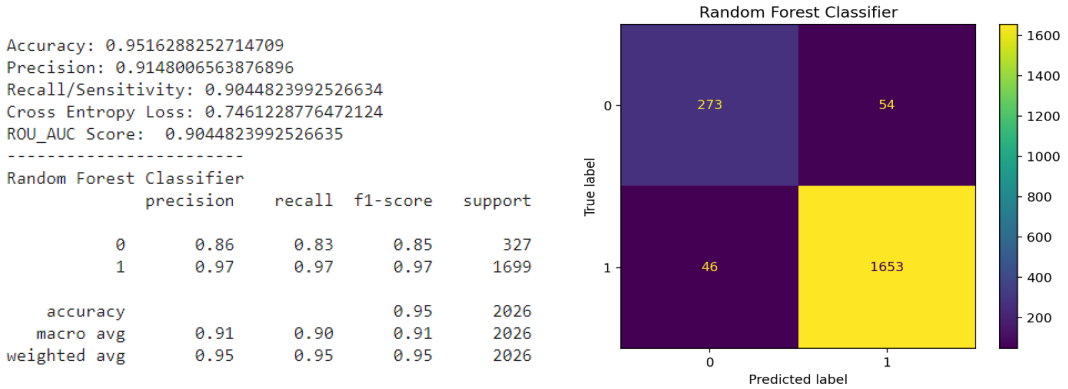
Hình 4.10: Performance Logistic Regression Model

- K-Neighbor Classifier: Mô hình phân loại K-Neighbor, có tới 261 dữ liệu với trạng thái Existing Customers đã được dự đoán chính xác và tối đa 1472 dữ liệu với trạng thái Attrited Customers đã được dự đoán chính xác, nhưng mô hình này cũng đưa ra dự đoán lỗi của 293 dữ liệu. Dựa trên những kết quả này, hiệu suất của mô hình thu được Accuracy Score là 86%, Precision score là 74.5%, Recall score là 83% với Cross Entropy Loss là 2,23.



Hình 4.11: Performance K-Neighbor Classifier Model

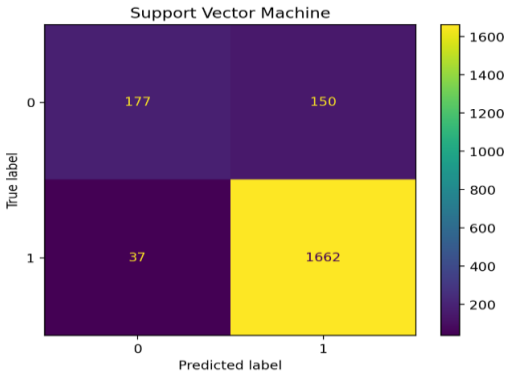
- Random Forest Classifier: Mô hình Phân loại rừng ngẫu nhiên, có tới 273 dữ liệu với trạng thái Existing Customers đã được dự đoán chính xác và tối đa 1653 dữ liệu với trạng thái Attrited Customers đã được dự đoán chính xác, nhưng mô hình này cũng đưa ra dự đoán lỗi của 100 dữ liệu. Dựa trên những kết quả này, hiệu suất của mô hình thu được Accuracy Score là 95%, Precision score là 91.5%, Recall score là 90% với Cross Entropy Loss là 0.76.



Hình 4.12: Performance Random Forest Classifier Model

- Support Vector Machine (SVM): Mô hình có tới 177 dữ liệu với trạng thái Existing Customers đã được dự đoán chính xác và tối đa 1662 dữ liệu với trạng thái Attrited Customers đã được dự đoán chính xác, nhưng mô hình này cũng đưa ra dự đoán lỗi của 187 dữ liệu. Dựa trên những kết quả này, hiệu suất của mô hình thu được Accuracy Score 91%, Precision score là 87%, Recall score là 76% với Cross Entropy Loss là 1.42.

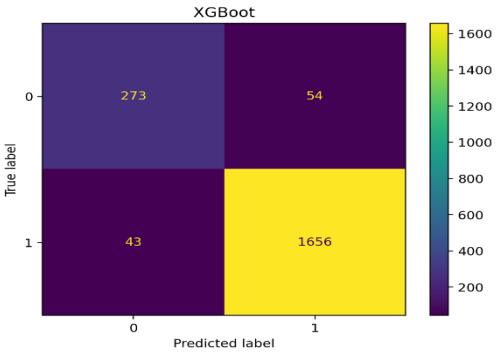
```
Accuracy: 0.9076999012833169
Precision: 0.8721606733923377
Recall/Sensitivity: 0.7597534437418665
Cross Entropy Loss: 1.4237242665370051
ROU_AUC Score:  0.7597534437418665
------------------------
Support Vector Machine
              precision    recall  f1-score   support

           0       0.83      0.54      0.65       327
           1       0.92      0.98      0.95      1699

    accuracy                           0.91      2026
   macro avg       0.87      0.76      0.80      2026
weighted avg       0.90      0.91      0.90      2026
```



Hình 4.13: Performance Support Vector Machine Model

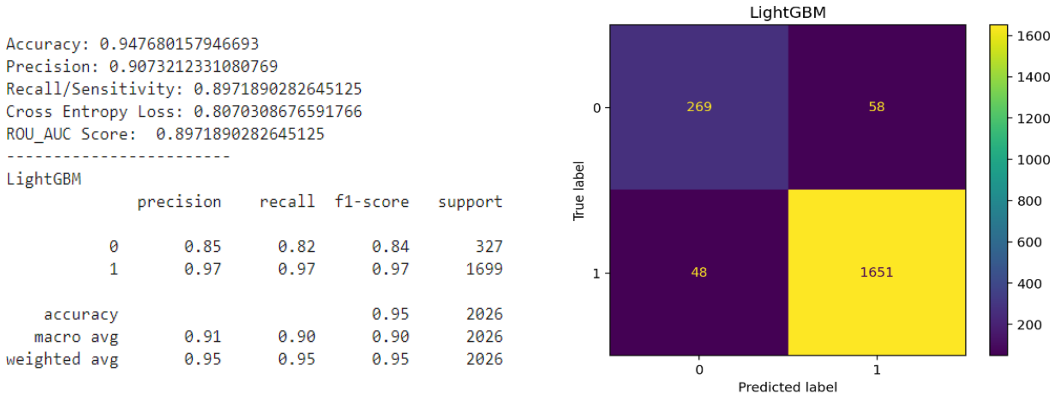
- Gradient Boosting Classifier: Mô hình Trình phân loại tăng cường độ dốc
  - Với giải thuật XGBoost có tới 273 dữ liệu với trạng thái Existing Customers đã được dự đoán chính xác và tối đa 1656 dữ liệu với trạng thái Attrited Customers đã được dự đoán chính xác, nhưng mô hình này cũng đưa ra dự đoán lỗi của 97 dữ liệu. Dựa trên những kết quả này, hiệu suất của mô hình thu được Accuracy Score 95%, Precision score là 91.6%, Recall score là 90.5% với Cross Entropy Loss là 0.74.

```
Accuracy: 0.9521224086870681
Precision: 0.9161725516322452
Recall/Sensitivity: 0.9047766900119336
Cross Entropy Loss: 0.7385093788957817
ROU_AUC Score:  0.9047766900119336
------------------------
XGBoot
              precision    recall  f1-score   support

           0       0.86      0.83      0.85       327
           1       0.97      0.97      0.97      1699

    accuracy                           0.95      2026
   macro avg       0.92      0.90      0.91      2026
weighted avg       0.95      0.95      0.95      2026
```



Hình 4.14: Performance XGBoost Model

- Với giải thuật LightGBM có tới 269 dữ liệu với trạng thái Existing Customers đã được dự đoán chính xác và tối đa 1651 dữ liệu với trạng thái Attrited Customers đã được dự đoán chính xác, nhưng mô hình này cũng đưa ra dự đoán lỗi của 106 dữ liệu. Dựa trên những kết quả này, hiệu suất của mô hình thu được Accuracy Score 95%, Precision score là 90.7%, Recall score là 89.7% với Cross Entropy Loss là 0.9.



Hình 4.15: Performance LightGBM Model

Dựa trên hiệu suất của các mô hình, có thể kết luận rằng mô hình Gradient Boosting Classifier với thuật toán XGBoost là mô hình hoạt động tốt nhất cho trường hợp tập dữ liệu này.

Model	Accuracy	Recall	Precision	F1	ROC_AUC
Random Forest Classifier	95.50839	91.27162	92.00139	97.32746	91.27162
XGBoost Classifier	95.21224	90.47767	91.61726	97.15459	90.47767
LightGBM Classifier	94.76802	89.7189	90.73212	96.88967	89.7189
KNeighbor Classifier	85.53801	83.22786	74.59616	90.94841	83.22786
Support Vector Machine	90.76999	75.97534	87.21607	94.67388	75.97534
Logistic regression	82.42843	72.7305	69.02548	89.25769	72.7305

Bảng 4.3: Thống kê hiệu suất của các mô hình với tham số mặc định

4.2.2. Điều chỉnh siêu tham số

Mỗi mô hình phải tuân theo các siêu tham số khác nhau cần được xác định. Tôi sử dụng các phương pháp random search và grid search methods để tối ưu hóa các siêu tham số cho các mô hình đã chọn của tôi.

- Grid search kiểm tra kỹ lưỡng mọi tổ hợp siêu tham số được cung cấp trong miền tham số để tối ưu hóa điểm số đã cho, đó là roc_auc trong dự án này.
- Random search kiểm tra các kết hợp ngẫu nhiên của siêu tham số cho một số lần lặp nhất định để tối ưu hóa điểm số đã cho. Ưu điểm của Random search là thời gian ngắn hơn nhưng sự đánh đổi là nó có thể bỏ lỡ sự kết hợp tốt nhất có thể.

4.2.2.1. Logistic Regression

Có một số siêu tham số mà chúng ta cần chú ý:

- C: Kiểm soát sức mạnh của regularization, regularization đang áp dụng một hình phạt đối với việc tăng cường độ của các giá trị tham số để giảm tình trạng thừa.
- Solver: Các thuật toán được sử dụng để tối ưu hóa.
- Class_weight: Trọng số được liên kết với các lớp.

Bằng Grid Search ta tìm được các tham số đưa kết quả mô hình tốt nhất với roc_auc score 0.92 và accuracy 85%.

C	class_weight	solver
0.4	balanced	lbfgs

Bảng 4.4: Tham số tối ưu hoá mô hình Logistic Regression

4.2.2.2. KNeighbors Classifier

Các siêu tham số quan trọng mà chúng ta sẽ điều chỉnh được hiển thị như sau:

- n_neighbors: số láng giềng gần nhất
- weights: hàm trọng số trong dự đoán
- p: Tham số công suất cho chỉ số Minkowski. Khi p = 1, điều này tương đương với việc sử dụng euclidean_ distance (l1) và euclidean_ distance (l2) cho p = 2

Bằng Grid Search ta tìm được các tham số đưa kết quả mô hình tốt nhất với roc_auc score 0.91 và accuracy 86%

n_neighbors	weights	p
100	distance	1

Bảng 4.5: Tham số tối ưu hoá mô hình KNeighbors Classifier

4.2.2.3. Random Forest Classifier

Dựa trên nghiên cứu của chúng tôi, có một số siêu tham số mà chúng tôi sẽ cần điều chỉnh:

- n_estimator: số lượng cây trong rừng
- min_samples_split: số lượng điểm dữ liệu tối thiểu cần có tại một nút lá
- max_features: số lượng tính năng cần xem xét khi tìm kiếm để phân chia tốt nhất
- max_depth: độ sâu tối đa của cây

Bằng Randomized Search ta tìm được các tham số đưa kết quả mô hình tốt nhất với roc_auc score 0.98 và accuracy 95%

n_estimator	min_samples_split	max_features	max_depth
700	3	sqrt	35

Bảng 4.6: Tham số tối ưu hoá mô hình Random Forest Classifier

4.2.2.4. XGBoost Classifier

Các siêu tham số mà chúng ta cần tinh chỉnh như sau:

- learning_rate: tốc độ học của việc tăng cường độ dốc
- max_depth: độ sâu cây tối đa
- min_child_weight: tổng trọng số tối thiểu (hessian) cần thiết. Hữu ích để giảm overfitting
- reg_lambda: L2 regularization

Bằng Grid Search ta tìm được các tham số đưa kết quả mô hình tốt nhất với roc_auc score 0.99 và accuracy 97%

learning_rate	max_depth	min_child_weight	reg_lambda
0.3	3	4	0.7

Bảng 4.7: Tham số tối ưu hoá mô hình XGBoost Classifier

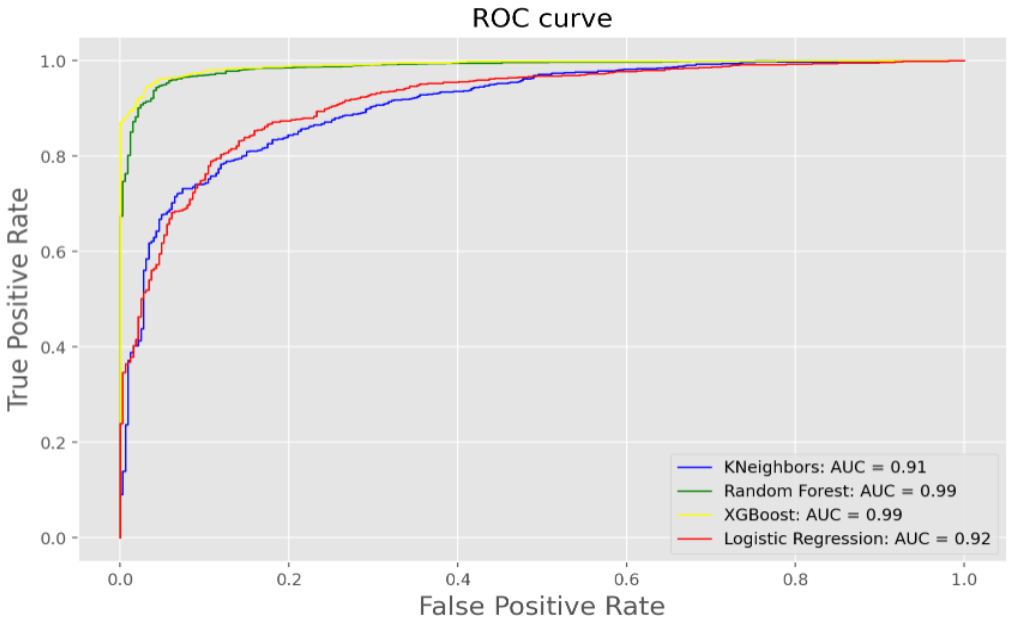
4.2.3. Tổng kết

Tôi tóm tắt các kết quả trong bảng và sơ đồ dưới đây. Có thể thấy XGBoost và Random Forest hoạt động tốt nhất trong số bốn mô hình về roc_auc. Đó là những mô hình dự đoán tốt nhất về vấn đề dữ liệu này.

Model	Accuracy	Recall	Precision	F1	ROC_AUC
XGBoost Classifier	96.54492	98.35197	97.54816	97.94842	99.13441
Random Forest Classifier	94.86673	98.76398	95.28677	96.99422	98.66237
Logistic Regression	84.79763	84.93231	96.52174	90.35692	91.62414
KNeighbor Classifier	86.47581	99.82343	86.22267	92.52591	90.98444

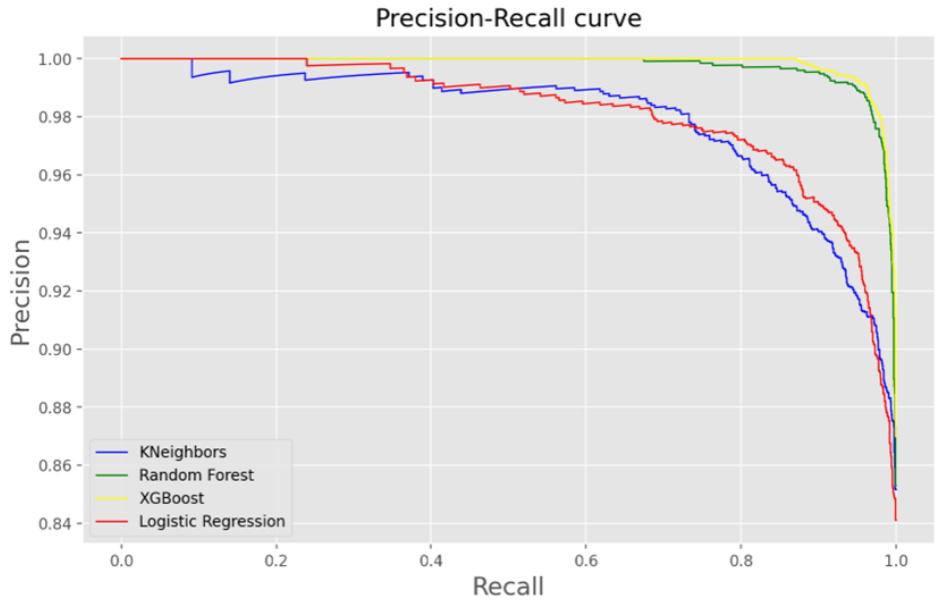
Bảng 4.8: Thống kê hiệu suất của các mô hình với tham số được tối ưu

Đường cong ROC là một công cụ quan trọng trong xác định tính hiệu quả của một phương pháp phân loại trong khoa học dữ liệu. Với sự trợ giúp của đường cong ROC, ta có thể đánh giá độ nhạy và độ đặc hiệu của một phương pháp phân loại cũng như tính toán diện tích dưới đường cong ROC để đánh giá hiệu quả của phương pháp. Ta có thể thấy XGBoost và Random Forest vẫn cho kết quả vượt trội.



Hình 4.16: ROC Curve

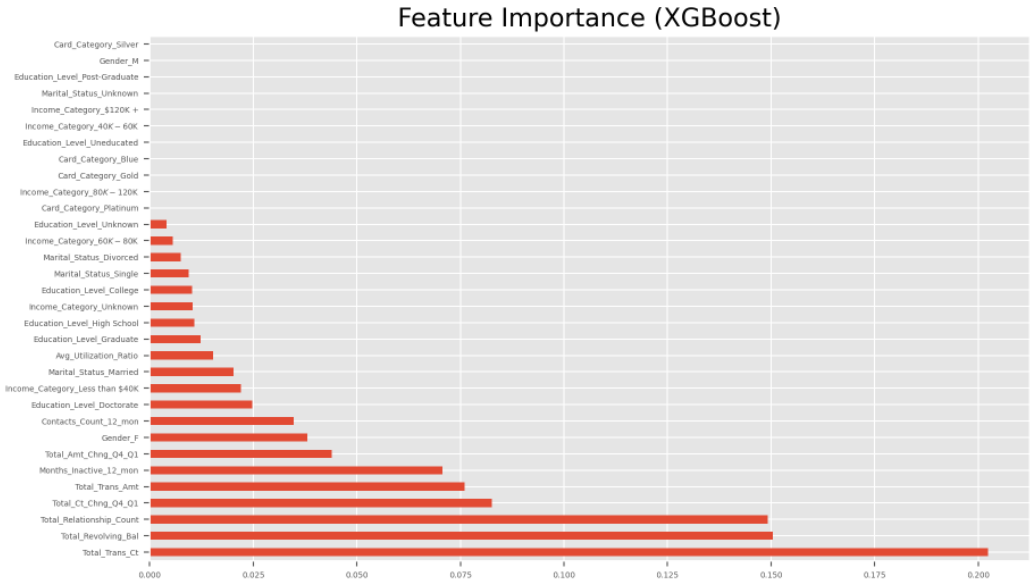
Chúng ta có thể nhìn vào Precision-Recall Curve để biết thêm thông tin. Nếu xét diện tích dưới đường cong, chúng ta có thể thấy XGBoost và Random Forest vẫn vượt trội so với các mô hình khác rất nhiều nhưng XGBoost có sự cải thiện rõ ràng hơn so với Random Forest.



Hình 4.17: Precision-Recall curve

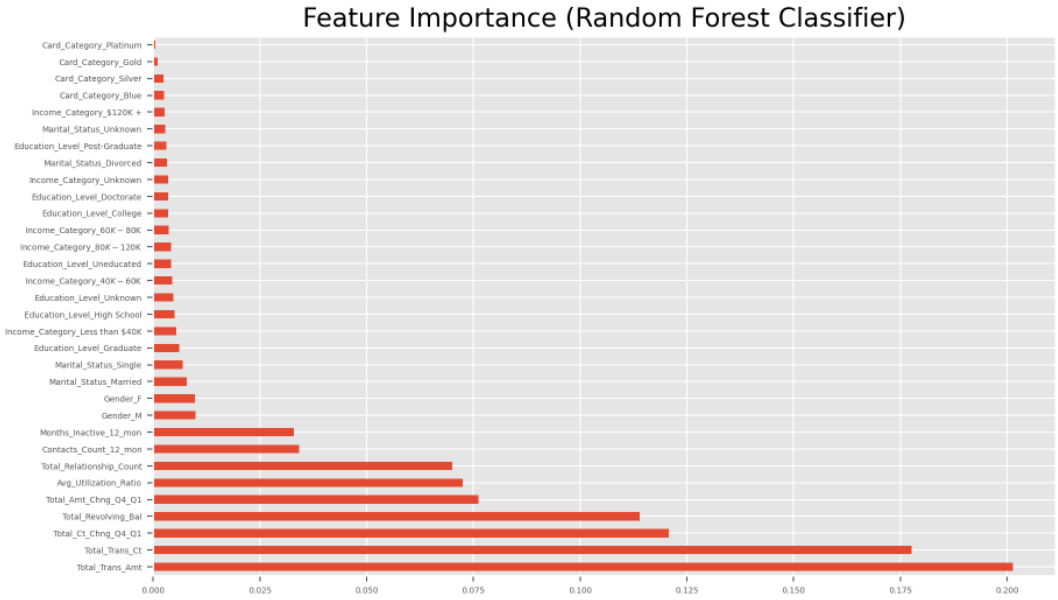
4.2.4. Tính năng quan trọng

- Top 3 tính năng quan trọng nhất từ mô hình XGBoost (theo thứ tự giảm dần):
    - Total_Trans_Ct (Tổng số giao dịch trong 12 tháng qua)
    - Total_Revolving_Bal (Tổng số dư quay vòng trên thẻ tín dụng)
    - Total_Relationship_Count (Tổng số sản phẩm do khách hàng nắm giữ)



Hình 4.17: Feature Importance XGBoost model

- Top 3 tính năng quan trọng hàng đầu từ mô hình Random Forest (theo thứ tự giảm dần):
    - Total_Trans_Amt (Tổng số tiền giao dịch trong 12 tháng qua)
    - Total_Trans_Ct (Tổng số lượng giao dịch trong 12 tháng qua)
    - Total_Ct_Chng_Q4_Q1 (Thay đổi về số lượng giao dịch từ Q4 thành Q1)



Hình 4.18: Feature Importance Random Forest model

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Bài luận này nhằm mục đích điều tra khả năng học máy để dự đoán tỷ lệ rời bỏ khách hàng thẻ tín dụng trong lĩnh vực ngân hàng. Tập dữ liệu được thu thập chứa hai loại dữ liệu: biến phân loại và biến liên tục. Những dữ liệu này mô tả 10.127 khách hàng trong ngân hàng; 1627 trong số những khách hàng này là khách hàng rời bỏ.

Năm mô hình học máy đã được đề xuất: Logistic regression, Random Forest, Support Vector Machine (SVM), KNeighbors, Gradient Boosting. Kết quả cho thấy XGBoost và Random Forest vượt trội so với các mô hình trước đó với ở các chỉ số hiệu suất khác nhau bao gồm: Recall, Precision, F1, ROC_AUC. Tổng số giao dịch, tổng số dư quay vòng trên thẻ tín dụng và sự thay đổi về số lượng giao dịch là ba biến số quan trọng hàng đầu để phát triển bất kỳ mô hình dự đoán khách hàng rời bỏ nào. Kết quả đã chứng minh rằng mô hình XGBoost có thể vượt trội so với các mô hình chức năng khác với độ chính xác accuracy lên đến 96.5%

5.2. Hướng phát triển

Trong tương lai, cần phân tích để hiểu rõ hơn về các biến độc lập tối ưu để phát triển một mô hình dự đoán rời bỏ mạnh mẽ hơn, chính xác hơn, nhanh hơn, ít phức tạp hơn và hiệu quả hơn. Một nghiên cứu như vậy sẽ bao gồm nhiều bộ dữ liệu hơn với các biến bổ sung để trích xuất các biến quan trọng nhất.

Ngoài ra, các mô hình học máy mới sẽ được thực thi để xác định đâu là mô hình tối ưu. Việc sử dụng mô hình học máy thông thường không phải lúc nào cũng đảm bảo kết quả tốt nhất cho một tập dữ liệu nhất định. Do đó, độ chính xác và hiệu quả của các mô hình dự đoán cần được cải thiện trong tương lai.

TÀI LIỆU THAM KHẢO

[1]. Harvard Business Review, The Value of Keeping the Right Customers, 2014 [https://hbr.org/2014/10/the-value-of-keeping-the-right-customers]

[2]. Bastan, M.; Bagheri Mazrae, M.; Ahmadvand, A. Dynamics of banking soundness based on CAMELS Rating system, 2016 [https://proceedings.systemdynamics.org/2016/proceed/papers/P1137.pdf]

[3]. Risselada, H.; Verhoef, P.C.; Bijmolt, T.H. Staying power of churn prediction models. J. Interact. Mark. 2010 [https://www.sciencedirect.com/science/article/abs/pii/S1094996810000253]

[4]. Domingos, E.; Ojeme, B.; Daramola, O. Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. Computation 2021 [https://www.mdpi.com/2079-3197/9/3/34]

[5]. Hadden, J.; Tiwari, A.; Roy, R.; Ruta, D. Computer assisted customer churn management: State-of-the-art and future trends. Comput. Oper. Res. 2007 [https://www.mdpi.com/2079-3197/9/3/34]

[6]. "IBM," [Online]. Available: [https://www.ibm.com/topics/machine-learning]

[7]. R. W. Picard, E. Vyzas and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," IEEE transactions on pattern analysis and machine intelligence, vol. 23, pp. 1175-1191, 2001.

[8]. Abraham Parangi , Co-Founder, CEO, Akkio 2023 Predicting Credit Card Customer Churn [https://www.akkio.com/post/credit-card-churn-prediction]

[9]. Xia, G.; He, Q. The research of online shopping customer churn prediction based on integrated learning. In Proceedings of the 2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018) [https://www.atlantis-press.com/proceedings/mecae-18/25893766]

[10]. Olaniyi, A.S.; Olaolu, A.M.; Jimada-Ojuolape, B.; Kayode, S.Y. Customer churn prediction in banking industry using K-means and support vector machine algorithms. Int. J. Multidiscip. Sci. Adv. Technol. 2020

[11]. Nie, G.; Rowe, W.; Zhang, L.; Tian, Y.; Shi, Y. Credit card churn forecasting by logistic regression and decision tree. Expert Syst. Appl. 2011 [https://www.sciencedirect.com/science/article/abs/pii/S0957417411009237?via%3Dihub]

[12]. Seng, J.L.; Chen, T.C. An analytic approach to select data mining for business decision. Expert Syst. Appl. 2010 [https://www.sciencedirect.com/science/article/abs/pii/S095741741000494X?via%3Dihub]

[13]. Kaya, E.; Dong, X.; Suhara, Y.; Balcisoy, S.; Bozkaya, B. Behavioral attributes and financial churn prediction. EPJ Data Sci. 2018 [https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0165-5]

[14]. Miao, X.; Wang, H. Customer churn prediction on credit card services using random forest method. In Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022), Online [https://www.atlantis-press.com/proceedings/icfied-22/125971597]

[15]. de Lima Lemos, R.A.; Silva, T.C.; Tabak, B.M. Propension to customer churn in a financial institution: A machine learning approach. Neural Comput. Appl. 2022 [https://link.springer.com/article/10.1007/s00521-022-07067-x]

[16]. Alfaiz, N.S.; Fati, S.M. Enhanced credit card fraud detection model using machine learning. Electronics 2022, 11, 662. [https://www.mdpi.com/2079-9292/11/4/662]

[17]. Machine Learning to Develop Credit Card Customer Churn Prediction, Dana AL-Najjar, Nadia Al-Rousan and Hazem AL-Najjar [https://www.mdpi.com/0718-1876/17/4/77#B12-jtaer-17-00077]

[18]. Overcoming Class Imbalance using SMOTE Techniques [https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/]

[19]. What is a Random Forest? [https://www.tibco.com/reference-center/what-is-a-random-forest]

[20]. SVM for churn prediction [https://subscription.packtpub.com/book/all-products/9781789345070/3/ch03lvl1sec24/svm-for-churn-prediction]

[21]. M. A. Hassonah, A. Rodan, A. -K. Al-Tamimi and J. Alsakran, "Churn Prediction: A Comparative Study Using KNN and Decision Trees," 2019 Sixth HCT Information Technology Trends (ITT), Ras Al Khaimah, United Arab Emirates, 2019.

[22]. What is Gradient Boosting and how is it different from AdaBoost? [https://www.mygreatlearning.com/blog/gradient-boosting/]

[23]. Sai Nikhilesh Kasturi, XGBOOST vs LightGBM: Which algorithm wins the race !!!, 2019 [https://towardsdatascience.com/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d]

[24]. "LightGBM," Microsoft Corporation, 2022. [Online]. Available: [https://lightgbm.readthedocs.io/en/v3.3.2/]

[25]. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," Advances in neural information processing systems, vol. 30, 2017.

[26]. A. Sharma, "Medium," 2018. [Online]. Available: [https://towardsdatascience.com/what-makes-lightgbm-lightning-fasta27cf0d9785e.]

[27]. A. Kumar, "Linkedin," May 2022. [Online]. Available: [https://www.linkedin.com/pulse/xgboost-vs-lightgbm-ashik-kumar]

[28]. Abhishek Sharma - What makes LightGBM lightning fast? , 2018 [https://towardsdatascience.com/what-makes-lightgbm-lightning-fast-a27cf0d9785e]

[29]. J. Mahmood, G.-E. Mustafa and M. Ali, "Accurate estimation of tool wear levels during milling, drilling and turning operations by designing novel hyperparameter tuned models based on LightGBM and stacking," Measurement, vol. 190, 2022.

[30]. J. Xu, Y. Zhang and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Information sciences, vol. 507, pp. 772-794, 2020.

[31]. Y. Li, T. Bellotti and N. Adams, "Issues using logistic regression with class imbalance, with a case study from credit risk modelling," Foundations of Data Science

[32]. A. Luque, A. Carrasco, A. Mart and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognition, vol. 91, pp. 216-231, 2019.

[33]. J. Brownlee, "Machine Learning Mastery," August 2020. [Online]. Available: [https://machinelearningmastery.com/confusion-matrix-machine-learning/]

[34]. Kiprono Elijah Koech, Cross-Entropy Loss Function Oct 3, 2020 [https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e]

[35]. T. Fawcett, "An introduction to ROC analysis," Pattern recognition letters, vol. 27, pp. 861-874, 2006.

[36]. M. Vuk and T. Curk, "ROC curve, lift chart and calibration plot," Advances in methodology and Statistics, vol. 3\, pp. 89-108, 2006.

[37]. S. Leteurtre, A. Martinot, A. Duhamel, F. Proulx, B. Grandbastien, J. Cotting, R. Gottesman, A. Joffe, J. Pfenninger and P. Hubert, "Validation of the paediatric logistic organ dysfunction (PELOD) score: prospective, observational, multicentre study," The Lancet, vol. 362, pp. 192-197, 2003.