

Xây dựng và đánh giá hiệu quả các mô hình học máy dựa trên cây quyết định với bài toán dự báo khách hàng ngừng tham gia tín dụng

(Building and evaluating machine learning models based on decision trees for predicting customer churn)

Trinh Gia Huy - 20520556@gm.uit.edu.vn

Phạm Lê Dịu Ái - 20520368@gm.uit.edu.vn

Lương Nguyễn Thành Nhân - 20520667@gm.uit.edu.vn

Nguyễn Thị Thảo Hồng - 20520192@gm.uit.edu.vn

Hà Lê Hoài Trung - trunghlh@uit.edu.vn

Tóm tắt nội dung—Với sự phát triển không ngừng của Internet và hoạt động ngân hàng, ngày càng nhiều người sử dụng thẻ tín dụng – credit card, do đó việc giữ chân khách hàng nhằm duy trì tỉ suất lợi nhuận là rất quan trọng đối với nhiều ngân hàng khi mà việc cạnh tranh giữa các ngân hàng rất khốc liệt. Do đó, nhóm tác giả đề xuất sử dụng học máy vào việc dự đoán việc khách sẽ ngừng sử dụng tín dụng để hỗ trợ ngân hàng đưa ra quyết định hoặc có thể thay đổi chương trình, dịch vụ để phục vụ và giữ được lượng khách hàng trung thành.

Với ưu điểm như dễ hiểu và dễ cấu hình, hiệu quả cao nên các thuật toán học máy dựa trên cây quyết định rất phù hợp cho bài toán trên. Bài nghiên cứu sử dụng 3 thuật toán ID3, Random Forest, XGBoost và đồng thời sử dụng 3 kỹ thuật Oversampling, Undersampling, SMOTE để xử lý mất cân bằng dữ liệu nhằm lập so sánh hiệu quả giữa các thuật toán đại diện cho 3 loại thuật toán, đơn giản - tổng hợp - tăng cường.

Kết quả thực nghiệm cho thấy, dựa trên độ đo ROC-AUC, kết quả cho thấy thuật toán XGBoost có độ chính xác cao nhất với tỉ lệ khoảng 97% và kỹ thuật xử lý mất cân bằng dữ liệu Oversampling cho ra kết quả có độ chính xác cao hơn 2 kỹ thuật còn lại.

Index Terms—học máy, cây quyết định, ngân hàng, rủi ro rời bỏ, tín dụng, ID3, XGBoost, Random Forest, SMOTE, Oversampling, Undersampling, phân loại nhị phân, khai phá dữ liệu.

I. GIỚI THIỆU

Tín dụng là một khái niệm quan trọng và có tầm quan trọng to lớn trong thời đại hiện nay. Trong một nền kinh tế phát triển, hệ thống tín dụng đóng vai trò quan trọng trong việc tạo điều kiện thuận lợi cho hoạt động kinh tế và tài chính khi mà nó thúc đẩy chỉ số tiêu dùng[1]. Tín dụng giúp khuyến khích tiêu dùng, đầu tư và phát triển kinh tế thông qua việc cung cấp nguồn vốn và tài trợ cho các cá nhân, hộ gia đình, doanh nghiệp và các tổ chức.

Tầm quan trọng của tín dụng không chỉ nằm ở việc tạo điều kiện cho các hoạt động kinh tế, mà còn đóng vai trò quan trọng trong cuộc sống cá nhân của mỗi người. Tín dụng giúp mọi người có khả năng tiếp cận đến các sản phẩm và dịch vụ mà họ không thể mua trực tiếp bằng tiền mặt. Nhờ tín dụng, người dùng có thể mua nhà, mua ô tô, đầu tư vào giáo dục, và thực hiện các dự án cá nhân khác mà không cần phải tích luỹ một số lượng lớn tiền mặt trước.

Với sự phát triển không ngừng của Internet và hoạt động ngân hàng, ngày càng nhiều người sử dụng thẻ tín dụng – credit card[2], do đó việc giữ chân khách hàng nhằm duy trì tỉ suất lợi nhuận là quan trọng đối với nhiều ngân hàng khi mà việc cạnh tranh giữa các ngân hàng rất khốc liệt. Các ngân hàng nhận ra rằng việc tìm kiếm khách hàng mới sẽ tốn nhiều chi phí hơn là việc giữ chân các khách hàng đang có[3]. Vì vậy vấn đề khách hàng ngừng sử dụng thẻ tín dụng trở thành mối quan tâm của các ngân hàng[4].

Có rất nhiều nguyên nhân để khách có thể dừng tiếp tục các dịch vụ tín dụng, do đó cần phải có những phương pháp nghiên cứu chuyên nghiệp và nghiêm túc, mà quản trị quan hệ khách hàng - Customer Relationship Management (CRM) bằng việc áp dụng các phương pháp khai phá dữ liệu - data mining cũng đang được áp dụng phổ biến trong nhiều lĩnh vực chứ không chỉ mỗi lĩnh vực tín dụng [5]. Việc hiểu rõ các nguyên nhân, yếu tố chính sẽ ảnh hưởng quyết định đến hướng giải quyết của doanh nghiệp và kịp thời giữ chân khách hàng [6]. Do đó, tác giả đề xuất sử dụng học máy vào việc dự đoán khách sẽ ngừng sử dụng tín dụng để hỗ trợ ngân hàng đưa ra quyết định hoặc có thể thay đổi chương trình, dịch vụ để phục vụ và giữ được lượng khách hàng.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Tiềm năng ứng dụng cao của bài toán không chỉ dừng lại ở lĩnh vực ngân hàng mà còn ở nhiều lĩnh vực khác càng thúc đẩy nhiều nghiên cứu có giá trị và đa dạng các phương pháp được áp dụng. Các thuật toán học máy từ cơ bản đến học sâu cũng được áp dụng cho bài toán này.

Trong bài nghiên cứu này, nhóm sẽ sử dụng các thuật toán cây quyết định ID3, Random Forest, XGBoost để áp dụng vào bài toán dự đoán khách hàng ngừng sử dụng tín dụng. Các thuật toán cây quyết định ID3, Random Forest, XGBoost đã được sử dụng trong các bài nghiên cứu về việc dự báo khách hàng rời bỏ như thuật toán ID3 được sử dụng và so sánh với mô hình kết hợp ID3 và Logistic của tác giả Choudhari và Potev, thuật toán ID3 cho được kết quả độ chính xác lớn hơn 95%[7]. Random Forest đạt độ chính xác 88.63% trong dự báo khách hàng rời bỏ của Irfan Ullah và đồng sự [8], cao hơn so với các thuật toán khác trong bài nghiên cứu. Trong bài nghiên

cứu của Roweida Mohammed và các đồng sự, việc sử dụng phương pháp Oversampling giúp Random Forest đạt được kết quả dự báo khách hàng rời bỏ với độ chính xác 99.8%[9]. Đối với XGBoost, thuật toán này đã được Raja và đồng sự sử dụng cho tập dữ liệu IBM Watson để tìm ra thuộc tính nào sẽ ảnh hưởng đến quyết định khách hàng rời bỏ và XGBoost đã cho được kết quả với độ chính xác đạt 79.8%[10]. Nghiên cứu của nhóm tác giả Lawanni với dự báo khách hàng trong lĩnh vực viễn thông cũng sử dụng XGBoost và đạt được độ chính xác 80.7%[11].

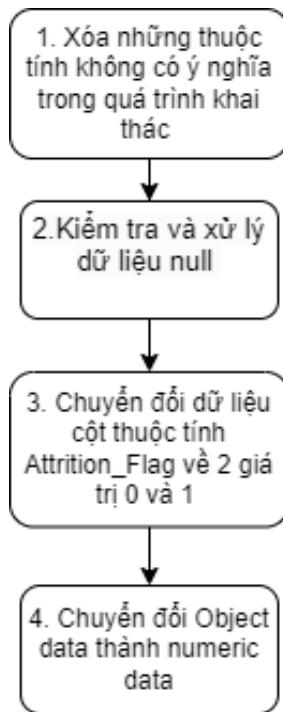
III. DỮ LIỆU

A. Tổng quan về dữ liệu

Bài nghiên cứu sử dụng bộ dữ liệu Credit Card customers được thu thập từ Kaggle¹ gồm 10127 dòng thông tin đặc trưng của từng khách hàng và 23 cột thuộc tính là các đặc trưng được thu thập như tình trạng hôn nhân, v.v... Với tỉ lệ khách hàng sẽ ngừng tham gia tín dụng là 16.07% trên toàn bộ dữ liệu, cho thấy dữ liệu bị mất cân bằng khá lớn, cần phải có những phương pháp xử lý dữ liệu thật sự hiệu quả nhằm đem đến một bộ dữ liệu đủ tốt để tăng độ hiệu quả hoạt động của các mô hình học máy.

B. Tiền xử lý dữ liệu

a) *Chuyển đổi dữ liệu:* Để thích hợp hơn cho việc phân tích và khai thác thì quá trình chuyển đổi dữ liệu nhằm đảm bảo tính chính xác, đáng tin cậy của dữ liệu và để thu được thông tin hữu ích từ tập dữ liệu. Đối với bộ dữ liệu này quy trình chuyển đổi dữ liệu như Hình 1.

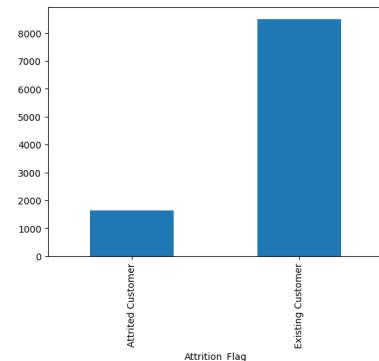


Hình 1. Mô hình chuyển đổi dữ liệu

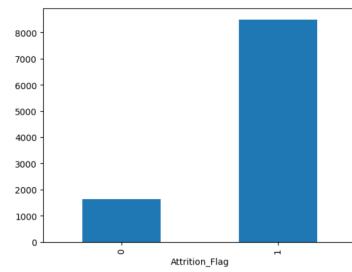
¹<https://www.kaggle.com/sakshigoyal7/credit-card-customers>

Thực hiện chuyển đổi các thuộc tính:

1) Thuộc tính Attrition_Flag được chuyển đổi từ dạng kiểu Categorical như hình 2 về dạng 2 giá trị 0 và 1 được mô tả như Hình 3.



Hình 2. Biểu đồ trước khi chuyển đổi của thuộc tính Attrition_Flag



Hình 3. Biểu đồ sau khi chuyển đổi của thuộc tính Attrition_Flag

2) Thuộc tính Gender, Card_Category, Marital_Status, Education_Level, Income_Category chuyển đổi về dạng số, đồng thời chuyển đổi kiểu dữ liệu cho các cột thuộc tính trên.

b) *Rời rạc hóa dữ liệu:* Nhằm chuyển đổi dữ liệu liên tục thành dữ liệu rời rạc bằng cách chia dữ liệu thành các khoảng hoặc các nhóm rời rạc thì quá trình này giúp thu gọn không gian dữ liệu và tạo ra các mẫu dữ liệu rời rạc. Việc rời rạc dữ liệu có thể được sử dụng để giảm độ phức tạp trong quá trình phân lớp dữ liệu. Quy trình rời rạc hóa dữ liệu phân chia các thuộc tính thành các lớp như Hình 4.

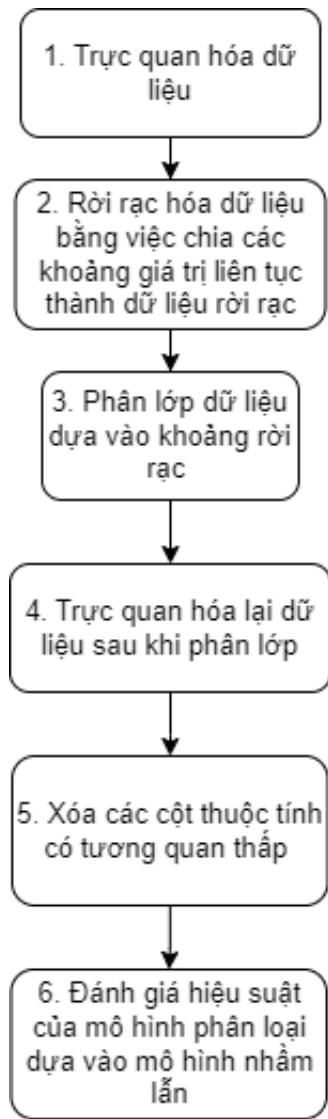
Thực hiện rời rạc hóa dữ liệu các thuộc tính lần lượt như sau:

1) Thuộc tính Customer_Age: Dữ liệu thuộc tính trải dài liên tục được biểu diễn như Hình 5, sau khi rời rạc hóa dữ liệu thay thế bằng các nhãn khái niệm số từ 1 đến 5 được biểu diễn biểu đồ như Hình 6.

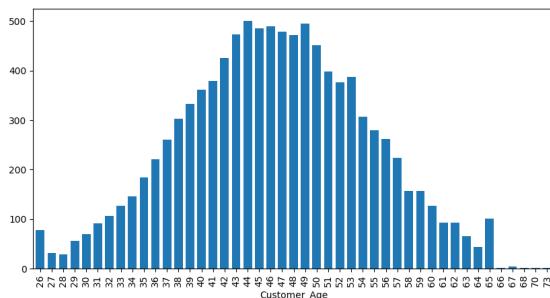
2) Thuộc tính Card_Category: Dữ liệu thuộc tính trải dài liên tục được biểu diễn như Hình 7, sau khi rời rạc hóa dữ liệu thay thế bằng các nhãn khái niệm Blue và Not Blue được biểu diễn biểu đồ như Hình 8.

3) Thuộc tính Months_on_book: Dữ liệu thuộc tính trải dài liên tục được biểu diễn như Hình 9, sau khi rời rạc hóa dữ liệu thay thế bằng các nhãn khái niệm số từ 1 đến 3 được biểu diễn biểu đồ như Hình 10.

4) Thuộc tính Credit_Limit: Dữ liệu thuộc tính trải dài liên tục được biểu diễn như Hình 11, sau khi rời rạc hóa dữ liệu



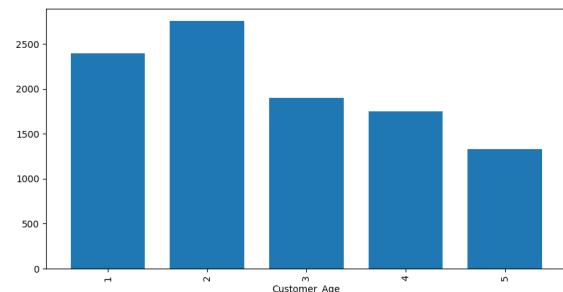
Hình 4. Mô hình rời rạc hóa dữ liệu



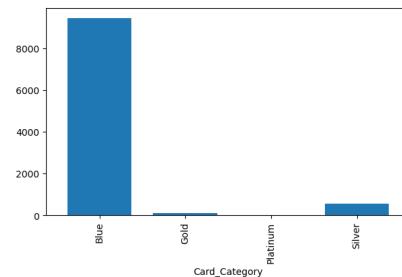
Hình 5. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Customer_Age

thay thế bằng các nhãn khái niệm số từ 1 đến 8 được biểu diễn biểu đồ như Hình 12.

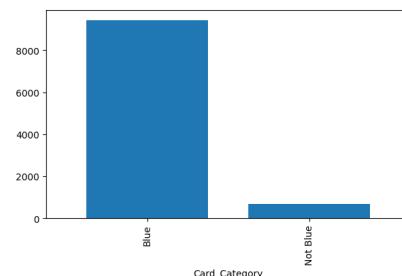
5) Thuộc tính Total_Revolving_Bal: Dữ liệu thuộc tính trái dài liên tục được biểu diễn như Hình 13, sau khi rời rạc hóa dữ liệu thay thế bằng các nhãn khái niệm số từ 1 đến 5 được biểu diễn biểu đồ như Hình 14.



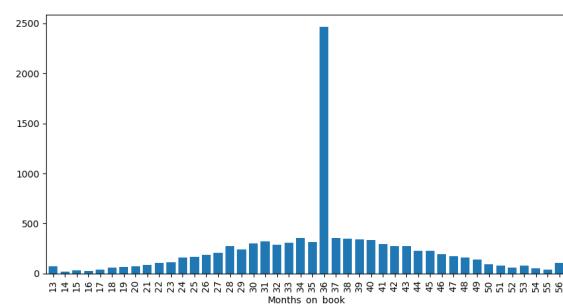
Hình 6. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Customer_Age



Hình 7. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Card_Category



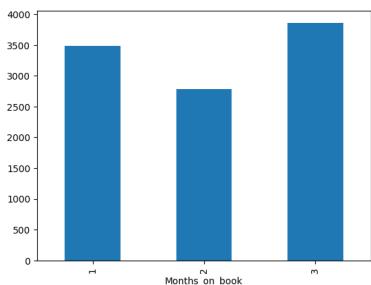
Hình 8. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Card_Category



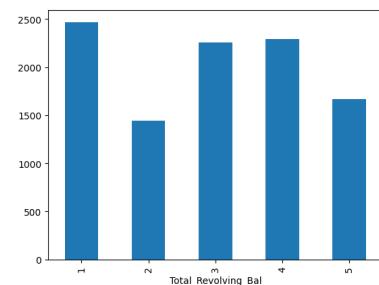
Hình 9. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Months_on_book

6) Thuộc tính Avg_Open_To_Buy: Dữ liệu thuộc tính trái dài liên tục được biểu diễn như Hình 15, sau khi rời rạc hóa dữ liệu thay thế bằng các nhãn khái niệm số từ 1 đến 5 được biểu diễn biểu đồ như Hình 16.

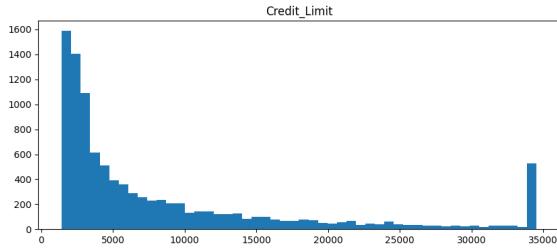
7) Thuộc tính Total_Amt_Chng_Q4_Q1: Dữ liệu thuộc tính trái dài liên tục được biểu diễn như Hình 17, sau khi rời rạc hóa dữ liệu thay thế bằng các nhãn khái niệm số từ 1 đến 5 được biểu diễn biểu đồ như Hình 18.



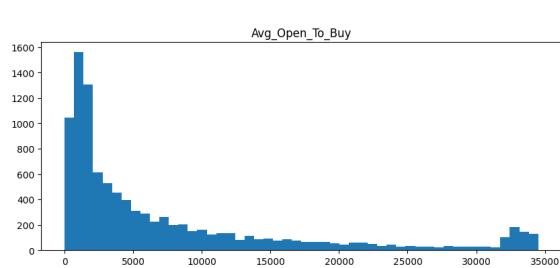
Hình 10. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Months_on_book



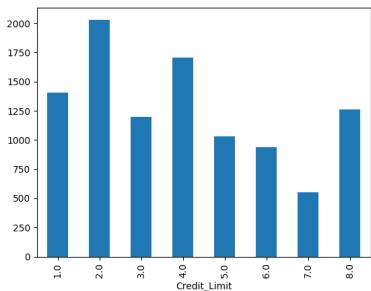
Hình 14. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Total_Revolving_Bal



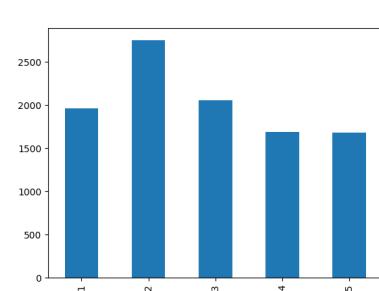
Hình 11. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Credit_Limit



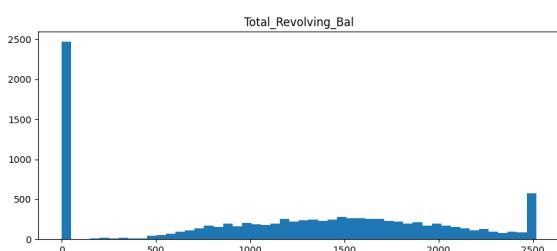
Hình 15. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Avg_Open_To_Buy



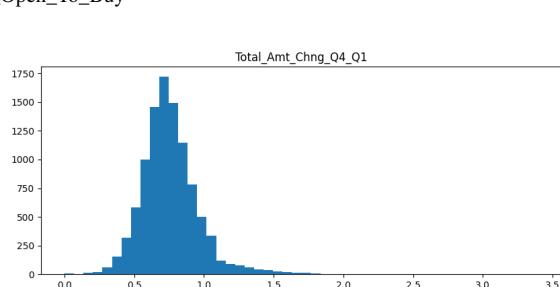
Hình 12. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Credit_Limit



Hình 16. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Avg_Open_To_Buy



Hình 13. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Total_Revolving_Bal



Hình 17. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Total_Amt_Chng_Q4_Q1

8) Thuộc tính Total_Trans_Amt: Dữ liệu thuộc tính trải dài liên tục được biểu diễn như Hình 19, sau khi rời rạc hóa dữ liệu thay thế bằng các nhãn khái niệm số từ 1 đến 5 được biểu diễn biểu đồ như Hình 20.

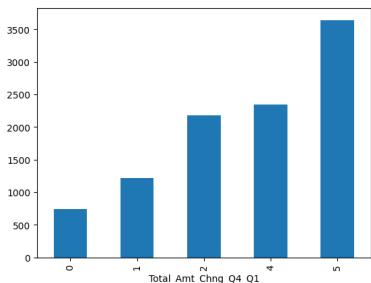
9) Thuộc tính Total_Trans_Ct: Dữ liệu thuộc tính trải dài liên tục được biểu diễn như Hình 21, sau khi rời rạc hóa dữ liệu thay thế bằng các nhãn khái niệm số từ 1 đến 5 được biểu diễn biểu đồ như Hình 22.

10) Thuộc tính Total_Ct_Chng_Q4_Q1: Dữ liệu thuộc tính trải dài liên tục được biểu diễn như Hình 23, sau khi rời rạc

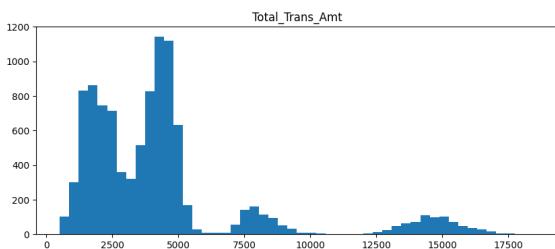
hóa dữ liệu thay thế bằng các nhãn khái niệm số từ 0 đến 5 được biểu diễn biểu đồ Hình 24.

11) Thuộc tính Avg_Utilization_Ratio: Dữ liệu thuộc tính trải dài liên tục được biểu diễn như Hình 25, sau khi rời rạc hóa dữ liệu thay thế bằng các nhãn khái niệm số từ 0 đến 4 được biểu diễn biểu đồ Hình 26.

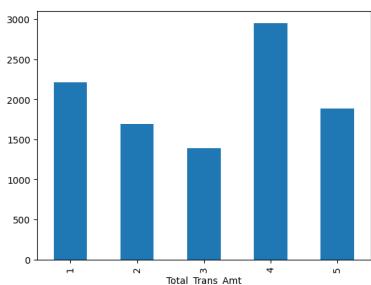
c) *Khai phá dữ liệu*: Là quá trình tìm ra các mẫu, quy luật, mối quan hệ hoặc sự kết hợp giữa các biến trong dữ liệu. Để có thể đánh giá hiệu suất cũng như đo lường mức độ chính



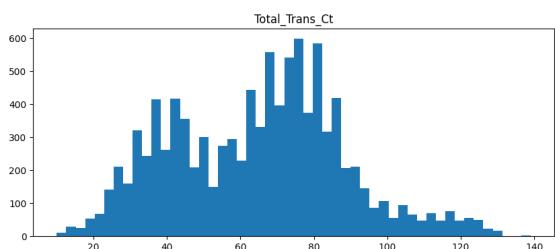
Hình 18. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Total_Amt_Chng_Q4_Q1



Hình 19. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Total_Trans_Amt



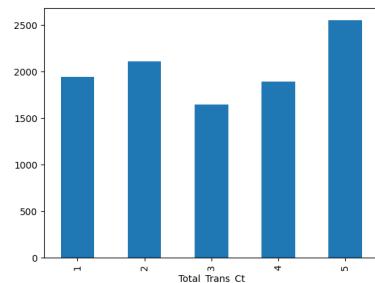
Hình 20. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Total_Trans_Amt



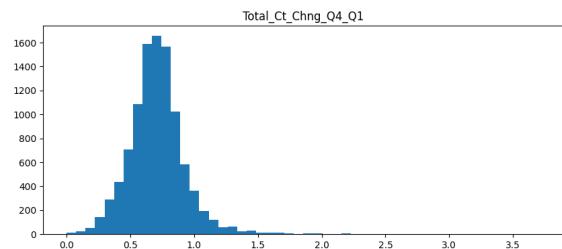
Hình 21. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Total_Trans_Ct

xác của mô hình phân loại sử dụng qua ma trận tương quan tuyến tính giữa các biến trong tập dữ liệu. Ma trận tương quan giữa các thuộc tính trong bộ dữ liệu được biểu diễn như Hình 27.

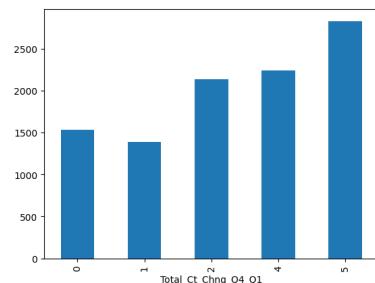
Loại bỏ các thuộc tính cho thấy mức độ tương quan thấp, không ảnh hưởng đến bộ dữ liệu như Card_Category, Education_Level, Marital_Status, Income_Category, Customer_Age. Sau khi loại bỏ các cột thuộc tính ta biểu diễn mức độ tương quan của các cột thuộc tính như Hình 28.



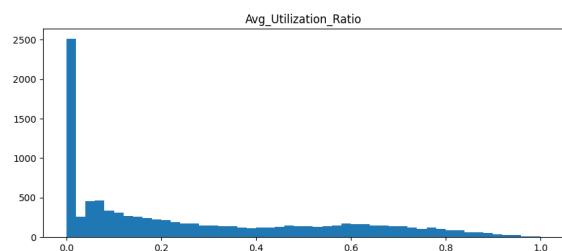
Hình 22. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Total_Trans_Ct



Hình 23. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Total_Ct_Chng_Q4_Q1



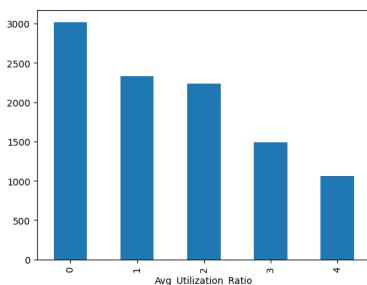
Hình 24. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Total_Ct_Chng_Q4_Q1



Hình 25. Biểu đồ biểu diễn trước khi rời rạc hóa dữ liệu thuộc tính Avg_Utilization_Ratio

d) **Mất cân bằng dữ liệu:** Làm việc với các bộ dữ liệu mất cân bằng là việc xảy ra rất phổ biến trong quá trình thực nghiệm các mô hình học máy, với bộ dữ liệu thực nghiệm cũng gặp vấn đề tương tự.

Để tăng hiệu quả huấn luyện trên bộ dữ liệu mất cân bằng, nhóm sử dụng phương pháp Oversampling (OS) và Undersampling(US) nhằm cải thiện mô hình. Được đánh giá là hiệu quả trong nhiều trường hợp, Oversampling là các phương pháp giúp giải quyết hiện tượng mất cân bằng mẫu bằng cách gia tăng kích thước mẫu thuộc nhóm thiểu số bằng các kĩ



Hình 26. Biểu đồ biểu diễn sau rời rạc hóa dữ liệu thuộc tính Avg_Utilization_Ratio

thuật khác nhau mà 2 kĩ thuật chính là lựa chọn mẫu có tái lập và mô phỏng mẫu mới dựa trên tổng hợp của các mẫu cũ. Mohammed và các đồng sự cũng đồng ý rằng [12], tuy nhiên điểm là kích thước mẫu sẽ bị giảm đáng kể, nhưng Undersampling là làm cân bằng mẫu một cách nhanh chóng, dễ dàng tiến hành thực hiện mà không cần đến thuật toán giả lập mẫu với cơ chế chính là giảm số lượng các quan sát của nhóm đa số để nó trở nên cân bằng với số quan sát của nhóm thiểu số. Tùy vào thực tế bài toán sử dụng, nếu ta biết khai thác điểm mạnh yếu của từng thuật toán thì có thể cải thiện tốt hiệu quả huấn luyện mô hình.

Bên cạnh 2 kĩ thuật trên, nhóm sẽ sử dụng thêm phương pháp SMOTE - Synthetic Minority Over-sampling, một phương pháp tự sinh tập mẫu để có thể khắc phục được tình trạng mất cân bằng dữ liệu. Được giới thiệu bởi N. V. Chawla và đồng sự vào 2001 [13], ý tưởng chung là nhằm gia tăng kích thước mẫu, với mỗi một mẫu thuộc nhóm thiểu số ta sẽ lựa chọn ra mẫu lỏng giềng gần nhất với nó và sau đó thực hiện tổ hợp tuyến tính để tạo ra mẫu giả lập. Phương pháp để lựa chọn ra các láng giềng của một quan sát có thể dựa trên thuật toán KNN hoặc SVM.

Bằng việc sử dụng cả 3 kĩ thuật, nhóm sẽ có thể lập một bảng đánh giá khách quan nhất về hiệu quả huấn luyện khi áp dụng các kĩ thuật trên và có thể kết luận được kĩ thuật nào là phù hợp nhất với từng thuật toán trên tập dữ liệu ban đầu.

e) *Tối ưu hóa siêu tham số - Tuning Hyper Parameter với Randomized Search:* Lựa chọn Hyperparameters chính xác cho các mô hình học máy và học sâu là một trong những cách tốt nhất để khai thác triệt để hiệu quả của các mô hình. Nhất là khi làm việc với các mô hình cây quyết định, việc sử dụng siêu tham số phù hợp sẽ giúp mô hình tránh việc dữ liệu quá khớp (overfitting) hoặc với dữ liệu mất cân bằng thì việc tinh chỉnh các siêu tham số cũng có thể giúp tinh chỉnh lại trọng số của từng lớp dữ liệu và khắc phục được tình trạng mất cân bằng. Trong bài nghiên cứu, nhóm sử dụng kĩ thuật Randomized Search từ thư viện Skicit-learn, một kĩ thuật được James Bergstra và Yoshua Bengio đánh giá là có nhiều ưu điểm hơn so với việc tinh chỉnh thủ công hay Grid search [14].

Cơ chế hoạt động của phương pháp Random Search bao gồm:

1) Xác định không gian siêu tham số: Xác định các siêu tham số cần tối ưu và phạm vi giá trị của chúng. Mỗi siêu tham số có một phạm vi giá trị cụ thể hoặc có thể được định nghĩa bằng cách sử dụng phân phối xác suất.

2) Tạo ra các bộ siêu tham số ngẫu nhiên: Tạo ra các bộ siêu tham số ngẫu nhiên từ không gian siêu tham số đã xác định. Mỗi bộ siêu tham số sẽ chứa một giá trị ngẫu nhiên cho mỗi siêu tham số.

3) Xây dựng và đánh giá mô hình: Với mỗi bộ siêu tham số, một mô hình học máy được xây dựng và huấn luyện bằng cách sử dụng tập dữ liệu huấn luyện. Sau đó, mô hình được đánh giá bằng cách sử dụng một phép đo hiệu suất như độ chính xác, F1-score, hoặc độ đo AUC-ROC.

4) Lưu giữ kết quả tốt nhất: Kết quả hiệu suất của mô hình được ghi lại cho mỗi bộ siêu tham số. Nếu mô hình hiện tại có kết quả tốt hơn so với các mô hình trước đó, thì nó sẽ trở thành mô hình tốt nhất cho đến thời điểm đó.

5) Lặp lại quá trình: Quá trình trên được lặp lại với một số lần xác định trước hoặc cho đến khi đạt được tiêu chí dừng như số lượng lần lặp hoặc đạt được kết quả đủ tốt.

6) Chọn mô hình tốt nhất: Sau khi quá trình tìm kiếm kết thúc, mô hình tốt nhất được chọn dựa trên kết quả hiệu suất đã được ghi lại.

IV. PHƯƠNG PHÁP LUẬN

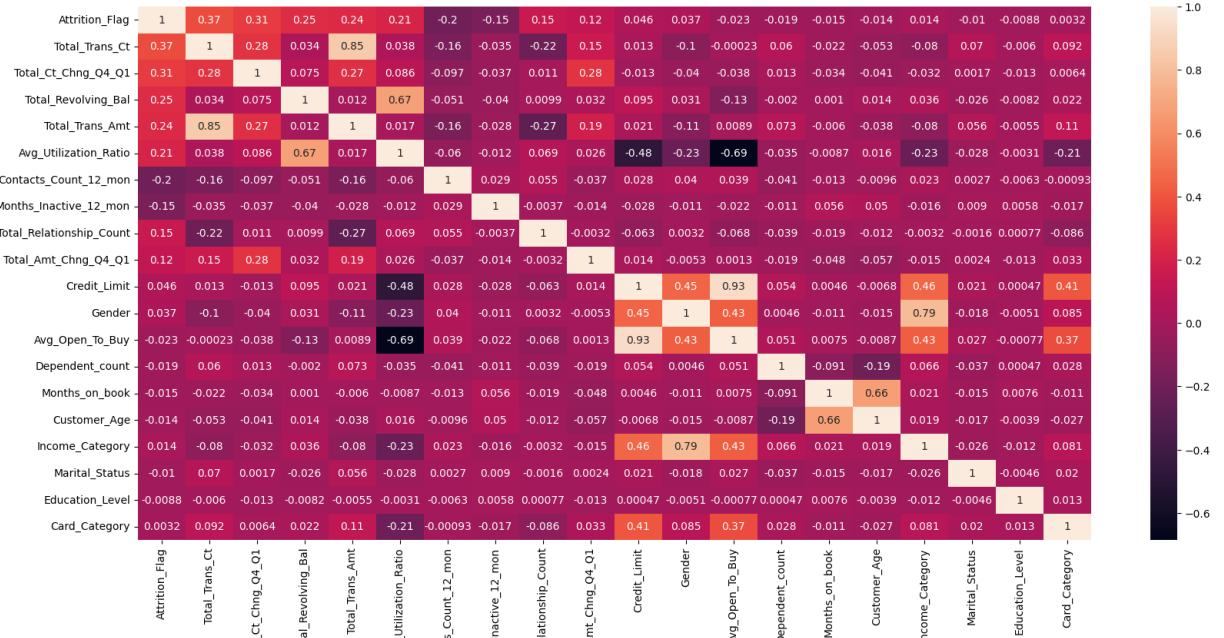
A. ID3

Được giới thiệu lần đầu tiên bởi Ross Quinlan vào năm 1986 [15] và là một trong những thuật toán đầu tiên được sử dụng để giải quyết bài toán phân loại trong học máy. Ý tưởng chính của thuật toán là xây dựng một cây nhị phân với mỗi nút trừ nút lá sẽ có 2 nhánh con tương ứng với 2 trường hợp xảy ra trong mô hình. Tùy vào từng loại bài toán và dạng dữ liệu sẽ có những cách xây dựng cây riêng biệt hơn. Với ý tưởng xây dựng cây như thế, nếu không có biện pháp hạn chế số nhánh hoặc độ sâu của cây quyết định, có khả năng cao gây tình trạng overfitting do lúc này các nút lá đều chỉ chứa dữ liệu của một lớp duy nhất.

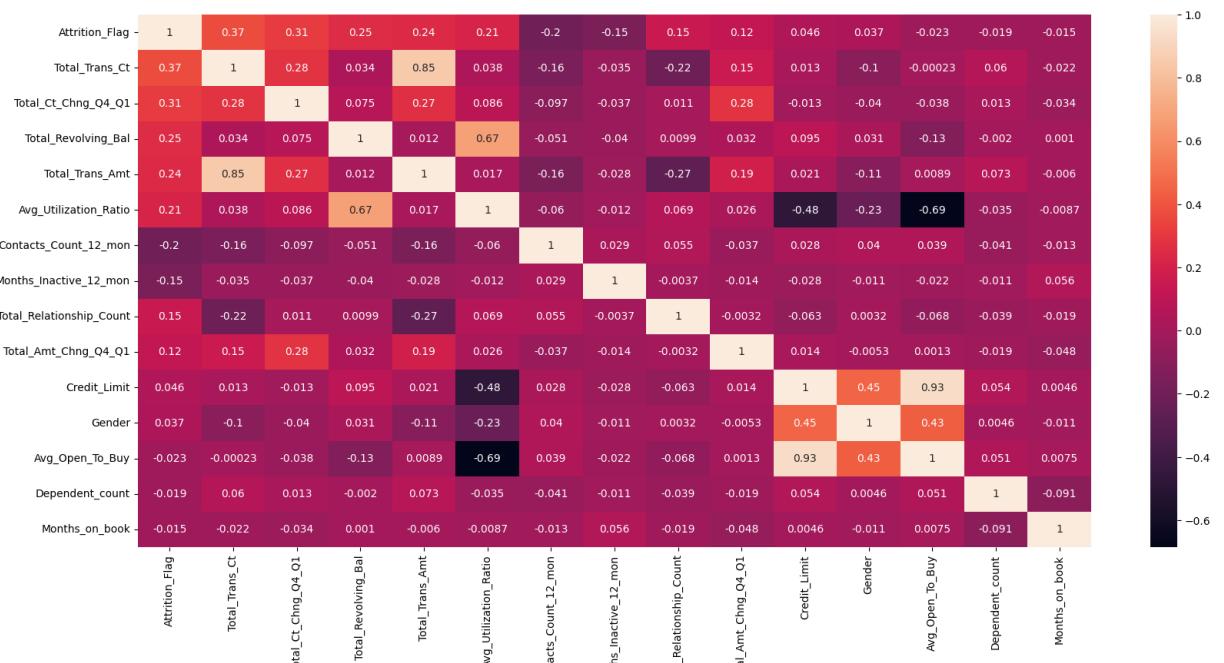
Trong ID3, chúng ta cần xác định thứ tự của thuộc tính cần được xem xét tại mỗi bước. Với các bài toán có nhiều thuộc tính và mỗi thuộc tính có nhiều giá trị khác nhau, việc tìm được nghiệm tối ưu thường là không khả thi. Thay vào đó, một phương pháp đơn giản thường được sử dụng là tại mỗi bước, một thuộc tính tốt nhất sẽ được chọn ra dựa trên một tiêu chuẩn nào đó (chúng ta sẽ bàn sớm). Với mỗi thuộc tính được chọn, ta chia dữ liệu vào các child node (nút con) tương ứng với các giá trị của thuộc tính đó rồi tiếp tục áp dụng phương pháp này cho mỗi nút con. Việc chọn ra thuộc tính tốt nhất ở mỗi bước như thế này được gọi là cách chọn greedy (tham lam). Cách chọn này có thể không phải là tối ưu, nhưng trực giác cho chúng ta thấy rằng cách làm này sẽ gần với cách làm tối ưu. Ngoài ra, cách làm này khiến cho bài toán cần giải quyết trở nên đơn giản hơn.

Sau mỗi câu hỏi, dữ liệu được phân chia vào từng child node tương ứng với các câu trả lời cho câu hỏi đó. Câu hỏi ở đây chính là một thuộc tính, câu trả lời chính là giá trị của thuộc tính đó. Để đánh giá chất lượng của một cách phân chia, chúng ta cần đi tìm một phép đo.

Trước hết, thế nào là một phép phân chia tốt? Bằng trực giác, một phép phân chia là tốt nhất nếu dữ liệu trong mỗi child node hoàn toàn thuộc vào một class—khi đó child node này có thể được coi là một leaf node, tức ta không cần phân



Hình 27. Ma trận tương quan tuyến tính ban đầu



Hình 28. Ma trận tương quan tuyến tính sau khi loại bỏ các thuộc tính tương quan thấp

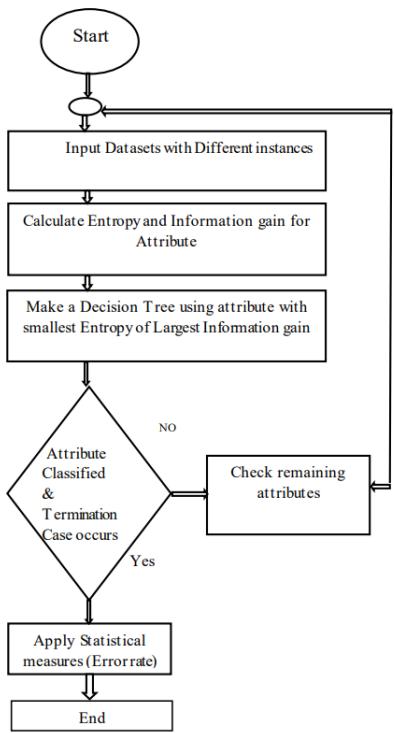
chia thêm nữa. Nếu dữ liệu trong các child node vẫn lẫn vào nhau theo tỉ lệ lớn, ta coi rằng phép phân chia đó chưa thực sự tốt. Từ nhận xét này, ta cần có một hàm số đo độ tinh khiết (purity), hoặc độ vẩn đục (impurity) của một phép phân chia. Hàm số này sẽ cho giá trị thấp nhất nếu dữ liệu trong mỗi child node nằm trong cùng một class (tinh khiết nhất), và cho giá trị cao nếu mỗi child node có chứa dữ liệu thuộc nhiều class khác nhau.

Một hàm số có các đặc điểm này và được dùng nhiều trong lý thuyết thông tin là hàm entropy.

Thuật toán ID3 có thể được mô hình hóa ý tưởng [16] như Hình 29:

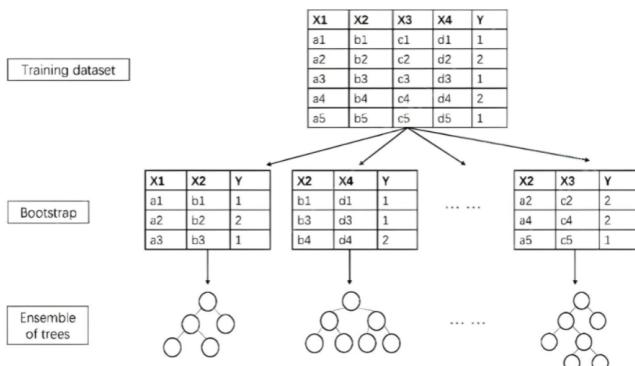
B. Random Forest

Random Forest là một thuật toán học máy phân loại và dự đoán được sử dụng rộng rãi trong các ứng dụng thực tế. Được



Hình 29. Mô hình hóa thuật toán ID3

giới thiệu lần đầu bởi Breiman (2001)[17], tác giả đã đề xuất thuật toán Random Forest được tạo nên từ tập hợp các cây quyết định được xây dựng dựa trên một bộ dữ liệu ngẫu nhiên của các mẫu dữ liệu và tập các thuộc tính quyết định cũng được lựa chọn ngẫu nhiên. Dữ liệu được của mỗi cây được lấy ngẫu nhiên và có thể trùng lặp để đảm bảo các kết quả dự đoán từ bộ dữ liệu huấn luyện khác biệt. Đối với bài toán phân lớp, kết quả dự đoán của rừng là dự đoán lặp lại nhiều nhất từ các cây quyết định. Còn đối với bài toán hồi quy, giá trị trung bình từ giá trị dự đoán của các cây quyết định là kết quả cuối cùng của rừng. Xem minh họa cấu trúc mô hình hóa thuật toán Random Forest ở Hình 30.



Hình 30. Mô hình hóa thuật toán Random Forest

Việc dự đoán từ kết quả của một tập các cây quyết định ngẫu nhiên với nhau nên giúp giảm tình trạng overfitting vốn thường xảy ra ở thuật toán cây quyết định và giúp mô hình có khả năng tổng quát hơn trong nhiều trường hợp, có khả năng

xử lý các tập dữ liệu lớn hơn và có độ phức tạp cao hơn, tuy nhiên cũng đòi hỏi tài nguyên tính toán cao hơn.

Random Forest không quy định chính xác một thuật toán cây quyết định nào sẽ được sử dụng để tạo nên rừng quyết định, nên tùy vào từng vấn đề bài toán (phân loại, hồi quy) cũng như đối với các kiểu dữ liệu khác nhau, có thể linh hoạt chọn các thuật toán cây quyết định phù hợp để lập rừng. Trong bài nghiên cứu, nhóm sử dụng thư viện Scikit-learn với các cây được tạo nên từ thuật toán cây quyết định Classification and Regression Trees - CART[18], đem lại cho bài nghiên cứu một so sánh tổng quát hơn khi áp dụng nhiều loại cây quyết định trong thực nghiệm.

Về cơ bản, Random Forest tổng quát được xây dựng như sau:

Đầu vào: Dataset (D) với N thuộc tính và n cây.

Đầu ra: Random Forest.

Lặp i=1 tới n:

Bước 1: Lấy bộ dữ liệu bootstrap (ngẫu nhiên) từ D.

Bước 2: Xây dựng rừng cây ngẫu nhiên từ bộ dữ liệu và lặp lại đến khi đạt được số node tối thiểu.

(1) Chọn 1 tập con của \sqrt{N} thuộc tính.

(2) Lặp từ j = 1 tới \sqrt{N} , chọn thuộc tính làm thuộc tính quyết định để chia cây tiếp tục thành 2 nhánh.

Giá trị dự đoán là giá trị xuất hiện nhiều nhất trong rừng.

C. XGBoost

Tianqi Chen và đồng sự đã giới thiệu Extreme Gradient Boosting - XGBoost vào năm 2016[19], đem lại một thuật toán gradient boosting tree-based (cây tăng cường độ dốc cực đại), được cải tiến dựa trên cây quyết định có khả năng giải quyết các bài toán phân loại với độ hiệu quả cao và khắc phục được nhiều nhược điểm vốn có của các thuật toán dựa trên cây quyết định.

Trong khi phương pháp AdaBoost - tổng hợp một mô hình dự đoán mạnh từ các mô hình yếu có ý tưởng là xây dựng một chuỗi các mô hình yếu tuần tự, trong đó mỗi mô hình cố gắng tập trung vào việc cải thiện phần của dữ liệu mà các mô hình trước đó dự đoán sai. Điều này đạt được bằng cách gán trọng số khác nhau cho từng mẫu dữ liệu trong quá trình huấn luyện. Thì XGBoost hoạt động bằng phương pháp Gradient Boosting - là sự kết hợp việc sử dụng một tập hợp các mô hình yếu (weak learner) để tạo ra một mô hình mới dự đoán mạnh hơn. Ý tưởng cơ bản của thuật toán gradient boosting là xây dựng một loạt các mô hình dự đoán tuần tự, trong đó mỗi mô hình mới cố gắng cải thiện sai số dự đoán của mô hình trước đó 31. Để làm điều này, thuật toán tối ưu hóa gradient boosting sử dụng phương pháp gradient descent.

Về cơ chế hoạt động:

1) Khởi tạo mô hình hóa thuật toán bằng một dự đoán đơn giản, thường là trung bình các giá trị mục tiêu trong tập huấn luyện.

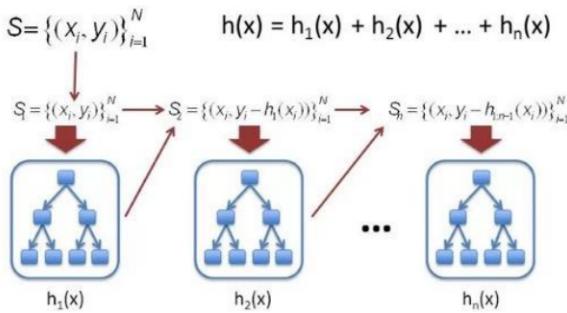
2) Tính toán sai số dự đoán giữa các dự đoán hiện tại và giá trị thực tế. Sai số này được gọi là gradient.

3) Một mô hình hóa thuật toán yếu mới được huấn luyện để dự đoán gradient của dữ liệu huấn luyện mô hình hóa thuật toán yếu này thường là một cây quyết định (decision tree) nhỏ, chỉ có một số lượng hạn chế các nút và độ sâu. Mô hình hóa

thuật toán yếu này được huấn luyện để xấp xỉ gradient của dữ liệu huấn luyện càng tốt càng tốt.

4) Khi mô hình hóa thuật toán yếu mới được huấn luyện, thuật toán cập nhật mô hình hóa thuật toán dự đoán bằng cách thêm một hệ số nhân với dự đoán của mô hình hóa thuật toán yếu. Hệ số này được tìm bằng cách tối ưu hóa mục tiêu mất mát (loss function) của bài toán.

Bằng cách này, mô hình hóa thuật toán dự đoán được cải thiện để xấp xỉ giá trị thực tế. Xem minh họa cấu trúc XGBoost ở Hình 31



Hình 31. Mô hình hóa thuật toán XGBoost

Về cơ bản, XGBoost được tính toán tổng quát như sau:

$l(x_1, x_2)$ - cost function

$f_i(x)$: thứ tự tại cây đang thực thi

$\hat{y}_i^0 = 0$: khởi tạo giá trị dự đoán ban đầu.

Giá trị đóng góp của cây thứ nhất trong việc dự đoán giá trị trong việc dự đoán giá trị trong mỗi mẫu dữ liệu (x_i)

$$\hat{y}_i^1 = \hat{y}_i^0 + f_1(x_i)$$

$$\hat{y}_i^2 = f_1(x_i) + f_2(x_i) = \hat{y}_i^1 + f_2(x_i)$$

Quá trình lặp đi lặp lại cho đến khi đạt được giá trị dự đoán cuối cùng.

$$\hat{y}_i^t = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

$$obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \text{sum}_{i=1}^t \Omega(f_i)$$

$$= \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{constant}$$

V. THỰC NGHIỆM

A. Thông số đánh giá

a) *Ma trận nhầm lẫn - Confusion Matrix*: Ma trận nhầm lẫn là đồ thị tương quan giữa 2 yếu tố Thực tế và Dự đoán, bao gồm 4 trường hợp, xem thêm ở Hình 32:

True Positive (TP): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Positive (dự đoán đúng)

True Negative (TN): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Negative (dự đoán đúng)

False Positive (FP): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Positive (dự đoán sai)

False Negative (FN): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Negative (dự đoán sai)

b) *ROC và AUC*: ROC là đường cong biểu diễn khả năng phân loại của một mô hình phân loại tại các ngưỡng threshold. Đường cong này dựa trên hai chỉ số :

1) TPR (true positive rate) chỉ số này sẽ đánh giá mức độ dự báo chính xác của mô hình trên positive. Khi giá trị của nó càng cao, mô hình dự báo càng tốt trên nhóm positive.

		Actual Class	
		1	0
Predicted Class	1	True Positive	False Positive
	0	False Negative	True Negative

Hình 32. Ma trận nhầm lẫn

2) FPR (false positive rate) là tỷ lệ dự báo sai các trường hợp thực tế là negative thành positive trên tổng số các trường hợp thực tế là negative. Một mô hình có FPR càng thấp thì mô hình càng chuẩn xác vì sai số của nó trên nhóm negative càng thấp.

AUC (Area Under The Curve) là chỉ số được tính toán dựa trên đường cong ROC (Receiving Operating Curve) nhằm đánh giá khả năng phân loại của mô hình.

c) *Accuracy*: accuracy hay độ chính xác được định nghĩa là tỷ lệ phần trăm dự đoán đúng cho dữ liệu thử nghiệm, được tính toán bằng cách chia số lần dự đoán đúng cho tổng số lần dự đoán.

$$\text{Accuracy} = \frac{TP+TN}{Total}$$

Tuy nhiên với dữ liệu mất cân bằng, việc sử dụng Accuracy làm thang đánh giá hiệu quả của mô hình sẽ dẫn đến nhiều ngộ nhận về độ chính xác thực tế của mô hình.

d) *Precision*: Precision cho thấy mức độ chuẩn xác của mô hình đối với các dữ liệu được dự báo với nhãn là Positive.

$$\text{Precision} = \frac{TP}{TP+FP}$$

e) *F1*: Tỉ lệ F1 là trung bình điều hòa giữa precision và recall. Do đó nó đại diện hơn trong việc đánh giá độ chính xác trên đồng thời precision và recall. Với bộ dữ liệu mất cân bằng thì việc sử dụng F1 làm thang đánh giá của mô hình sẽ là tối ưu nhất.

$$F1 = \frac{2}{precision^{-1} + recall^{-1}}$$

f) *Recall*: Tỉ lệ Recall được định nghĩa là độ phủ của các dữ liệu được dự đoán thuộc về một lớp so với tất cả các dữ liệu thực sự thuộc về lớp đó.

$$\text{Recall} = \frac{TP}{TP+FN}$$

B. Quy trình

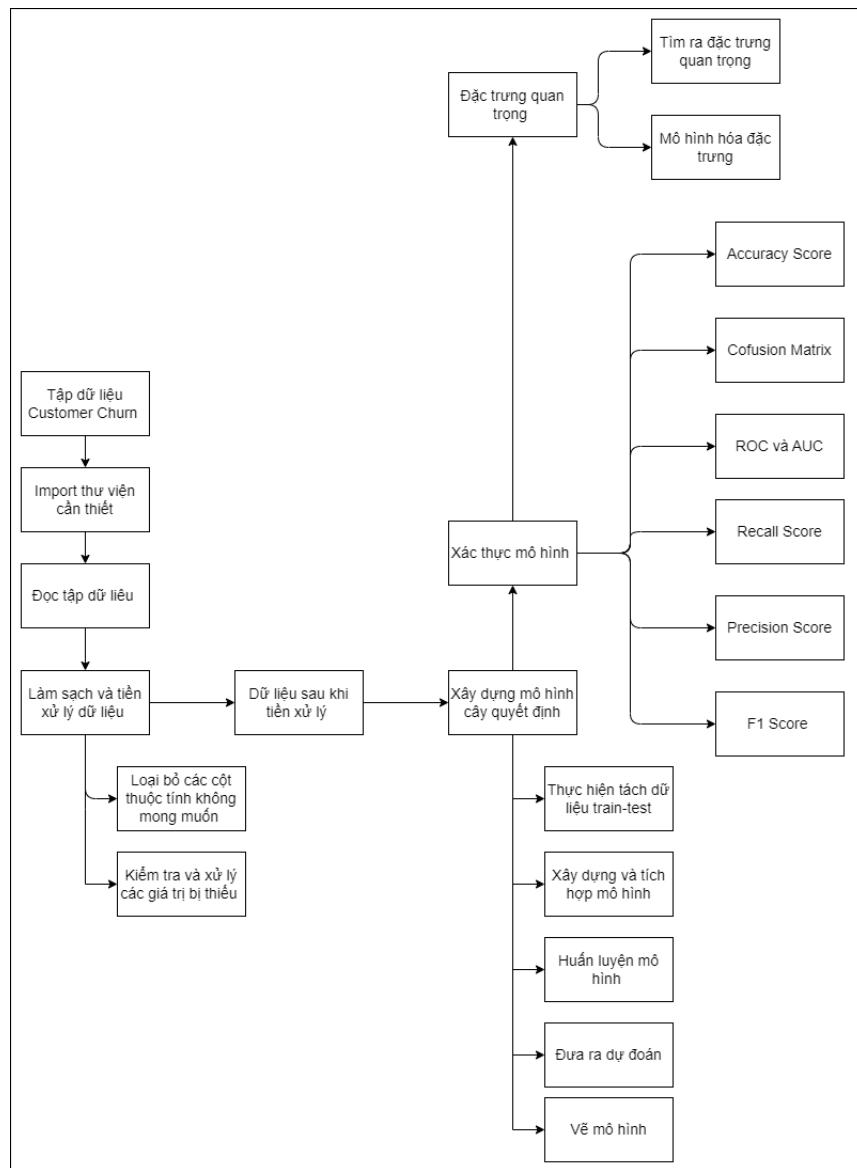
Quy trình thực nghiệm bao gồm các giai đoạn chính như Hình 33.

1) Thực hiện chia tập dữ liệu 7-3 và dùng kỹ thuật Oversampling để cân bằng dữ liệu theo kiểu Default value và tính toán các thông số đánh giá:

a) Thuật toán ID3

Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 34.

Biểu đồ biểu diễn cây quyết định như Hình 35.



Hình 33. Mô hình hóa quy trình thực nghiệm.

Đường cong ROC theo kiểu Default value được biểu diễn như Hình 36.

b) Thuật toán RF

Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 46.

Đường cong ROC theo kiểu Default value được biểu diễn như Hình 47.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng theo kiểu Default value như Hình 48.

c) Thuật toán XGBoost

Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 49.

Đường cong ROC theo kiểu Default value được biểu diễn như Hình 50.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 51.

2) Thực hiện chia tập dữ liệu 7-3 và dùng kỹ thuật Oversampling để cân bằng dữ liệu và kết hợp Randomized

Search để tính toán các thông số đánh giá:

a) Thuật toán ID3

Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 43.

Biểu đồ biểu diễn cây quyết định như Hình 44

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 45.

b) Thuật toán RF

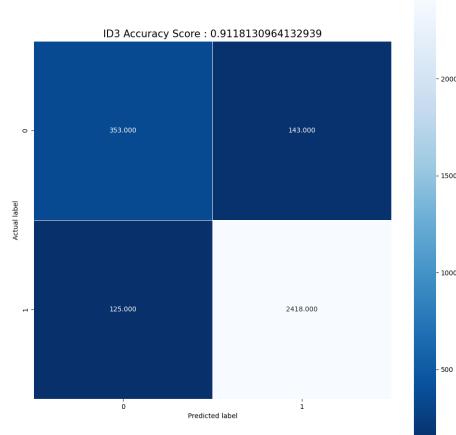
Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 46.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 47.

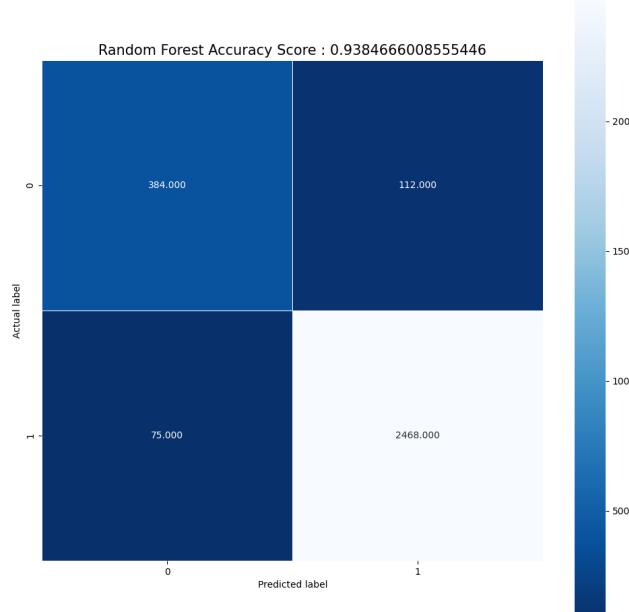
Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 48.

c) Thuật toán XGBoost

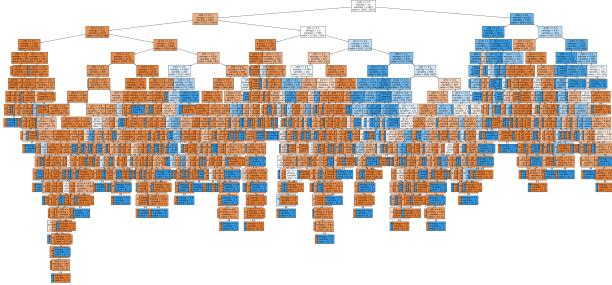
Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 49.



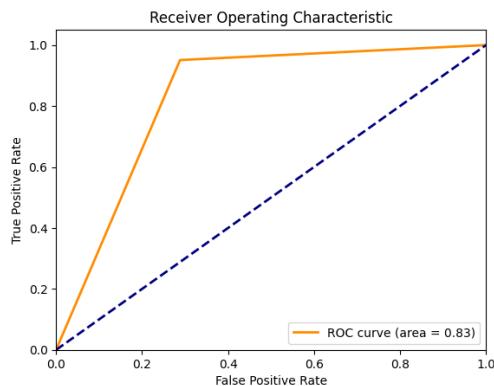
Hình 34. Biểu đồ biếu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật OS theo kiểu Default value.



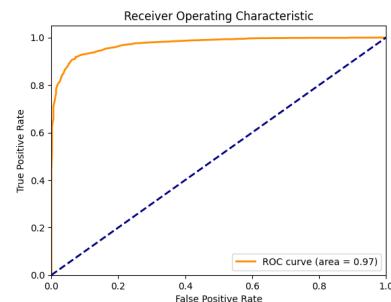
Hình 37. Biểu đồ biếu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật OS theo kiểu Default value.



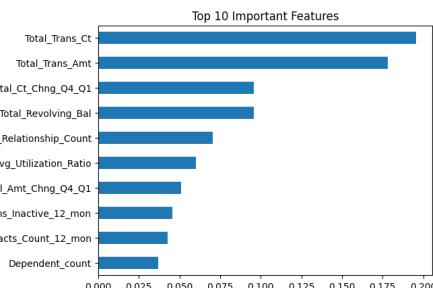
Hình 35. Biểu đồ biếu diễn cây quyết định khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật OS theo kiểu Default value.



Hình 36. Biểu đồ biếu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật OS theo kiểu Default value.



Hình 38. Biểu đồ biếu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật OS theo kiểu Default value.



Hình 39. Biểu đồ biếu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật OS theo kiểu Default value.

Đường cong ROC khi kết hợp với Randomized Search được biếu diễn như Hình 50.

Biểu đồ biếu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 51.

3) Thực hiện chia tập dữ liệu 7-3 và dùng kỹ thuật SMOTE để cân bằng dữ liệu theo kiểu Default value và tính toán các thông số đánh giá:

a) Thuật toán ID3

Thông số đánh giá Accuracy theo kiểu Default value được biếu diễn như Hình 52.

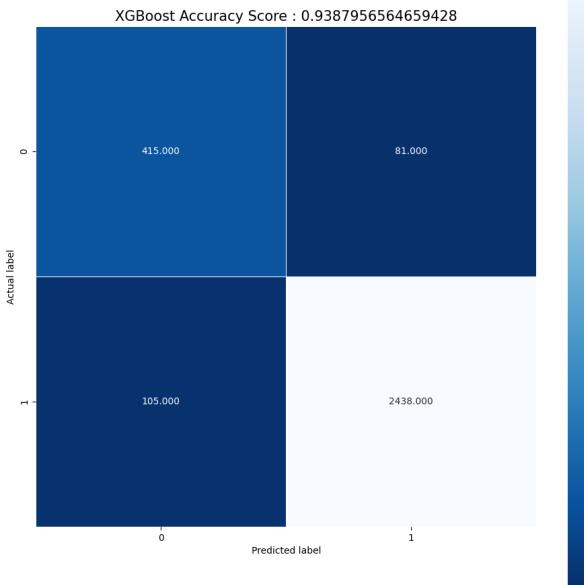
Biểu đồ biếu diễn cây quyết định như Hình 53

Đường cong ROC theo kiểu Default value được biếu diễn như Hình 54.

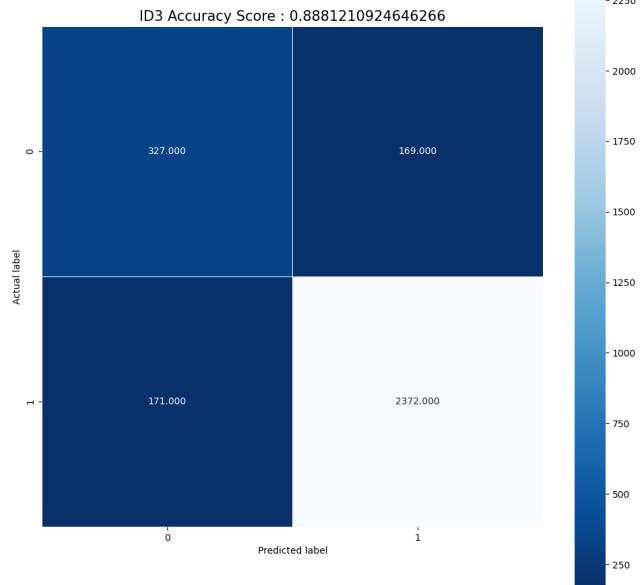
b) Thuật toán RF

Thông số đánh giá Accuracy theo kiểu Default value được biếu diễn như Hình 55.

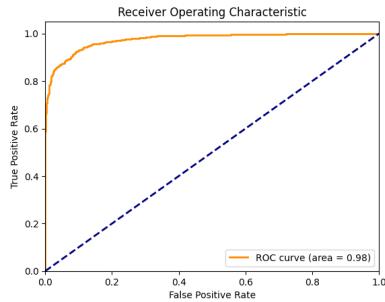
Đường cong ROC theo cách Randomized Search được biếu diễn như Hình 56.



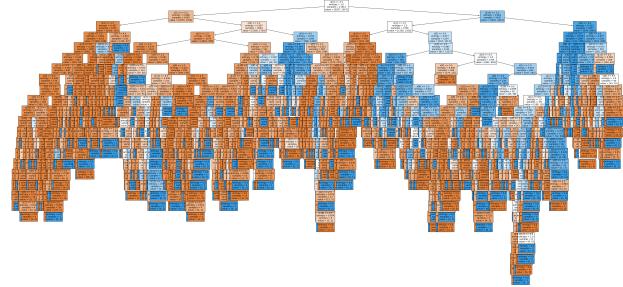
Hình 40. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật OS theo kiểu Default value.



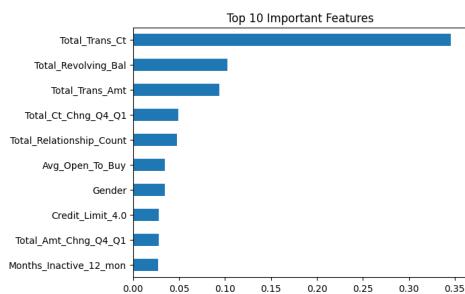
Hình 43. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật OS và Randomized Search.



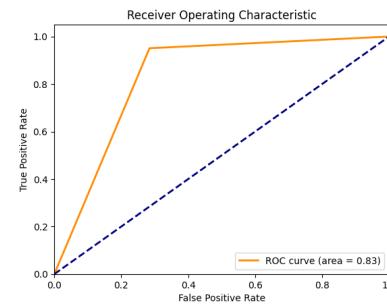
Hình 41. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật OS theo kiểu Default value.



Hình 44. Biểu đồ biểu diễn cây quyết định khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật OS và Randomized Search.



Hình 42. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật OS theo kiểu Default value.



Hình 45. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật OS và Randomized Search.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 57.

c) Thuật toán XGBoost

Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 58.

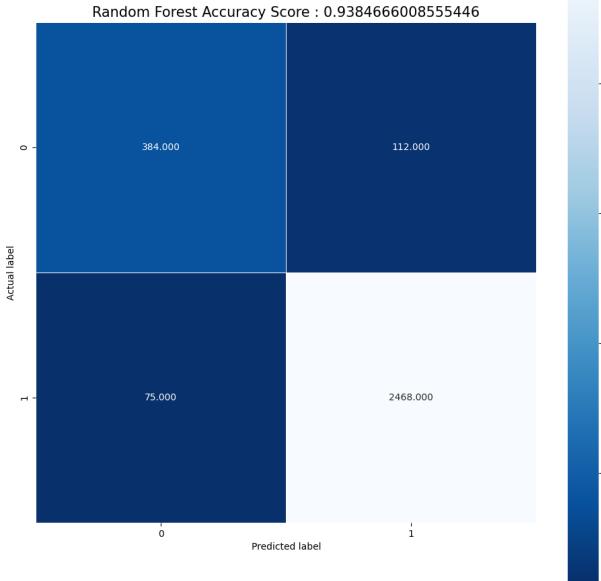
Đường cong ROC theo kiểu Default value được biểu diễn như Hình 59.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 60.

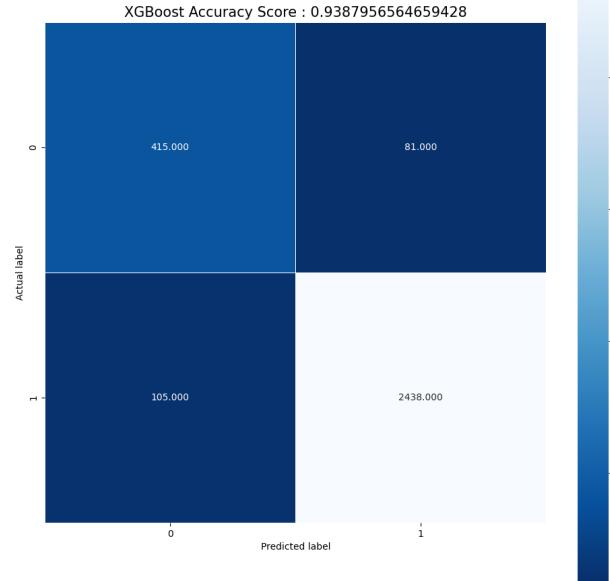
4) Thực hiện chia tập dữ liệu 7-3 và dùng kỹ thuật SMOTE để cân bằng dữ liệu và kết hợp Randomized Search để tính toán các thông số đánh giá:

a) Thuật toán ID3

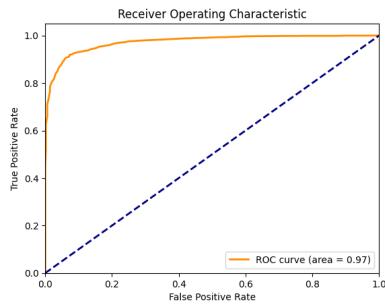
Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 61.



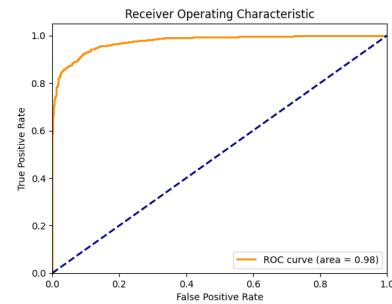
Hình 46. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật OS và Randomized Search.



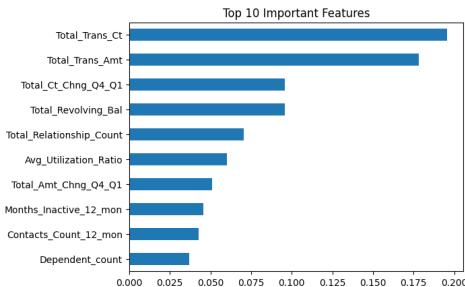
Hình 49. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật OS và Randomized Search.



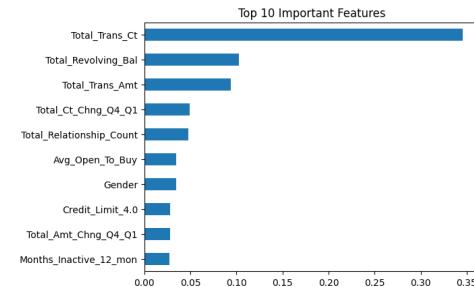
Hình 47. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật OS và Randomized Search.



Hình 50. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật OS và Randomized Search.



Hình 48. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật OS và Randomized Search.



Hình 51. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật OS và Randomized Search.

Biểu đồ biểu diễn cây quyết định như Hình 62.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 63.

b) Thuật toán RF

Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 64.

Đường cong ROC khi kết hợp với Randomized Search được

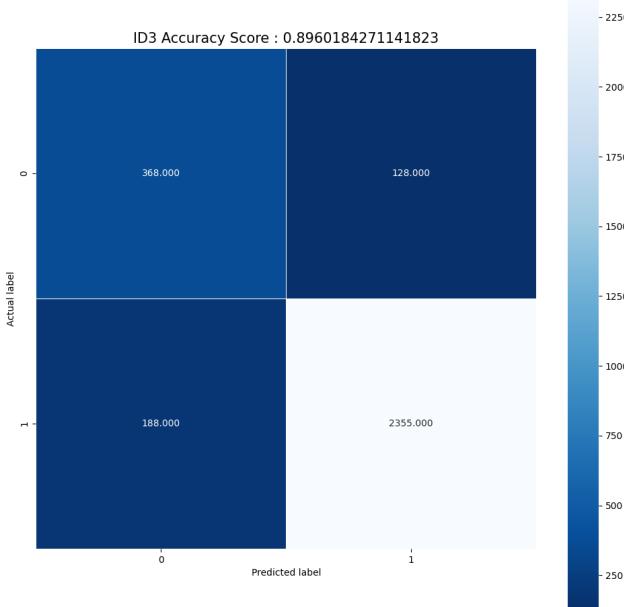
biểu diễn như Hình 65.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 66.

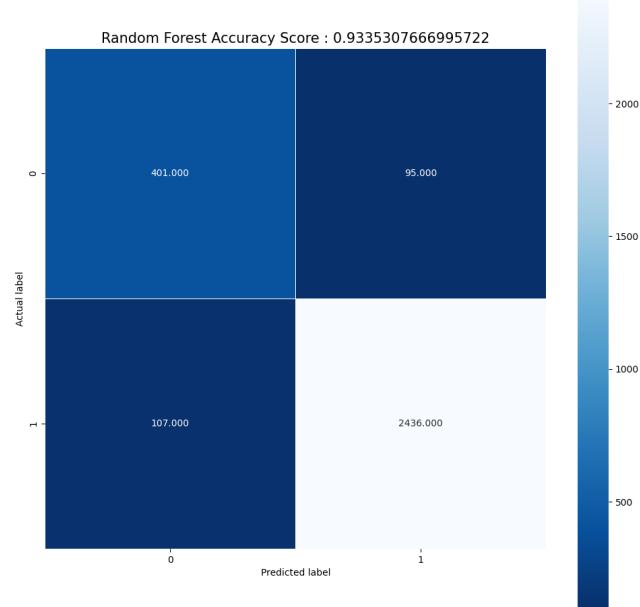
c) Thuật toán XGBoost

Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 67.

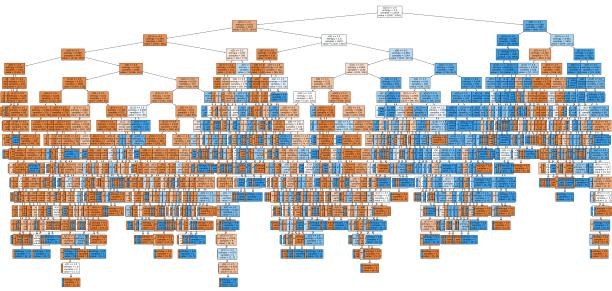
Đường cong ROC khi kết hợp với Randomized Search được



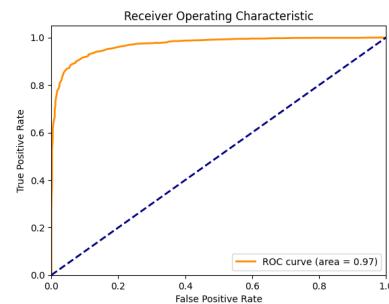
Hình 52. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật SMOTE theo kiểu Default value.



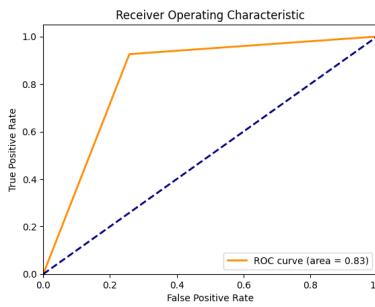
Hình 55. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật SMOTE theo kiểu Default value.



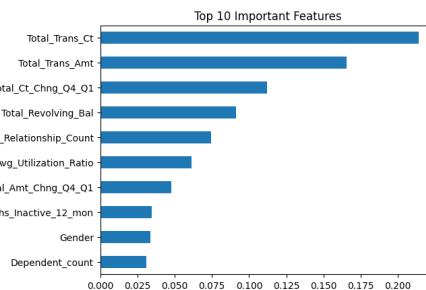
Hình 53. Biểu đồ biểu diễn cây quyết định khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật SMOTE theo kiểu Default value.



Hình 56. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật SMOTE theo kiểu Default value.



Hình 54. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật SMOTE theo kiểu Default value.



Hình 57. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật SMOTE theo kiểu Default value.

biểu diễn như Hình 68.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 69.

5) Thực hiện chia tập dữ liệu 7-3 và dùng kỹ thuật Undersampling để cân bằng dữ liệu theo kiểu Default value và tính toán các thông số đánh giá:

a) Thuật toán ID3

Thông số đánh giá Accuracy theo kiểu Default value được

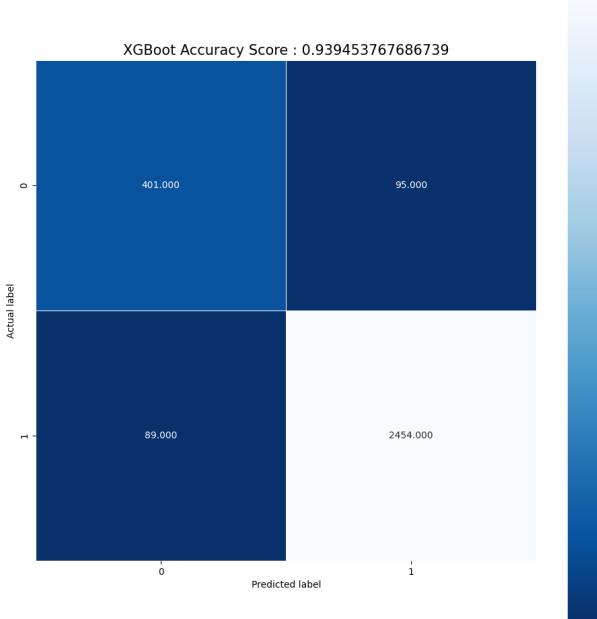
biểu diễn như Hình 70.

Biểu đồ biểu diễn cây quyết định như Hình 71.

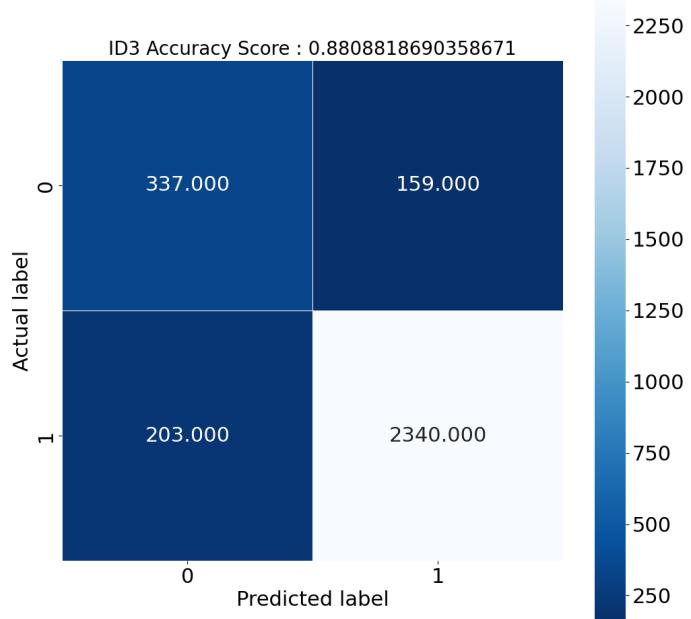
Đường cong ROC theo kiểu Default value được biểu diễn như Hình 72.

b) Thuật toán RF

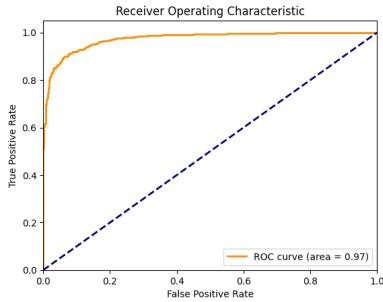
Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 73.



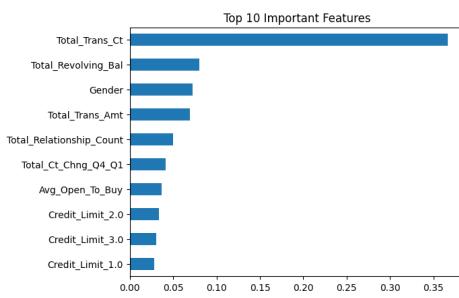
Hình 58. Biểu đồ biếu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật SMOTE theo kiểu Default value.



Hình 61. Biểu đồ biếu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật SMOTE và Randomized Search.



Hình 59. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật SMOTE theo kiểu Default value.



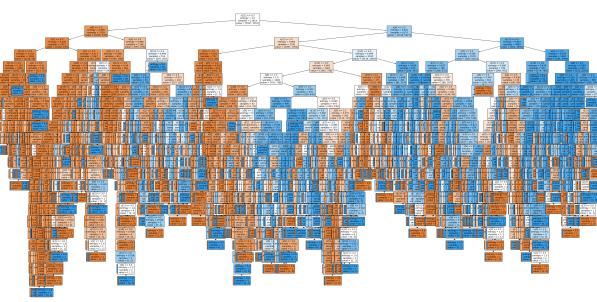
Hình 60. Biểu đồ biếu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật SMOTE theo kiểu Default value.

Đường cong ROC theo kiểu Default value được biếu diễn như Hình 74.

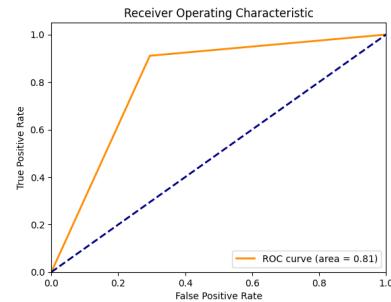
Biểu đồ biếu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 75.

c) Thuật toán XGBoost

Thông số đánh giá Accuracy theo kiểu Default value được



Hình 62. Biểu đồ biếu diễn cây quyết định khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật SMOTE và Randomized Search.



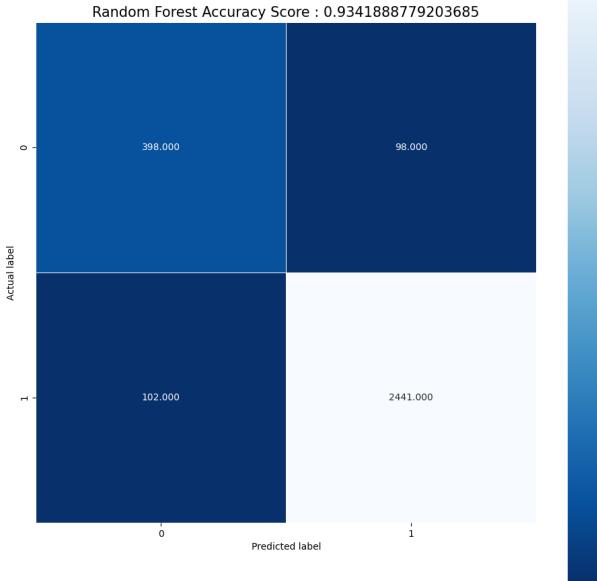
Hình 63. Biểu đồ biếu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật SMOTE và Randomized Search.

biểu diễn như Hình 76.

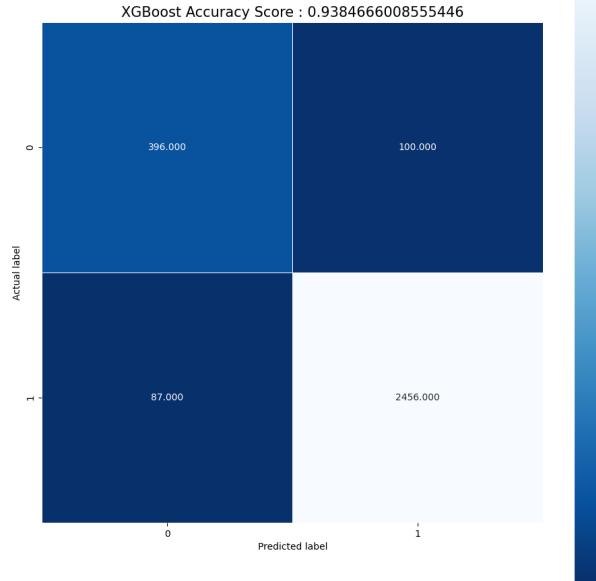
Đường cong ROC theo kiểu Default value được biếu diễn như Hình 77.

Biểu đồ biếu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 78.

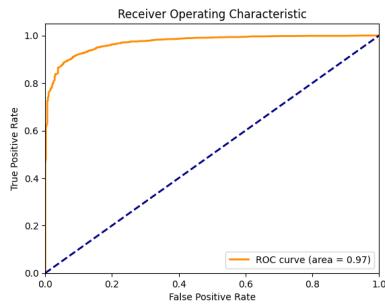
6) Thực hiện chia tập dữ liệu 7-3 và dùng kỹ thuật Undersampling để cân bằng dữ liệu và kết hợp Randomized



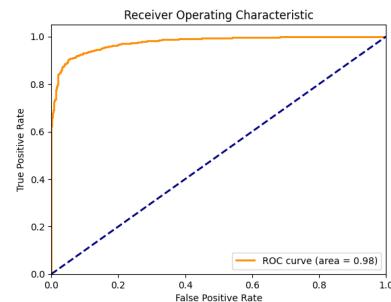
Hình 64. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật SMOTE và Randomized Search.



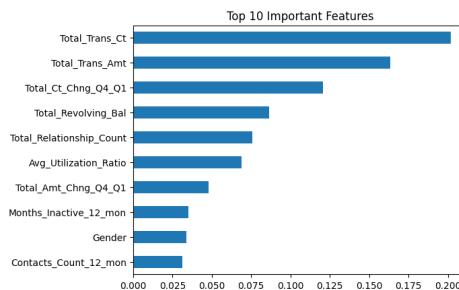
Hình 67. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật SMOTE và Randomized Search.



Hình 65. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật SMOTE và Randomized Search.



Hình 68. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật SMOTE và Randomized Search.



Hình 66. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật SMOTE và Randomized Search.

Search để tính toán các thông số đánh giá:

a) Thuật toán ID3

Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 79.

Biểu đồ biểu diễn cây quyết định như Hình 80

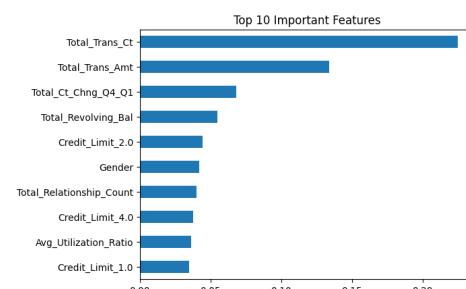
Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 81.

b) Thuật toán RF

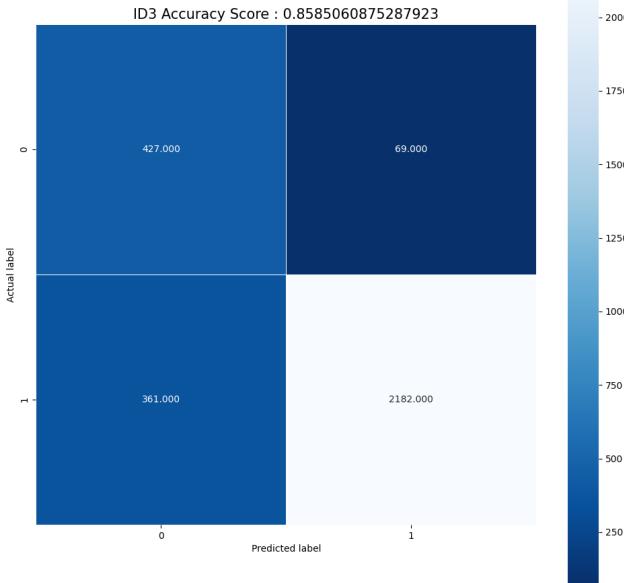
Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 82.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 83.

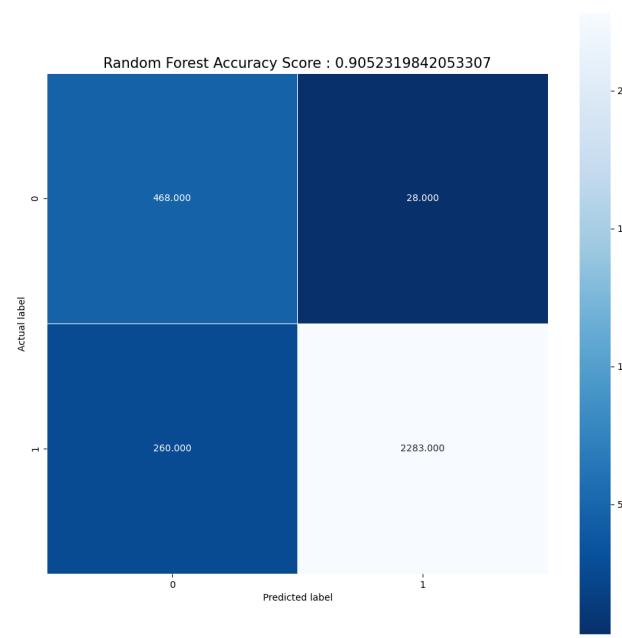
Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc



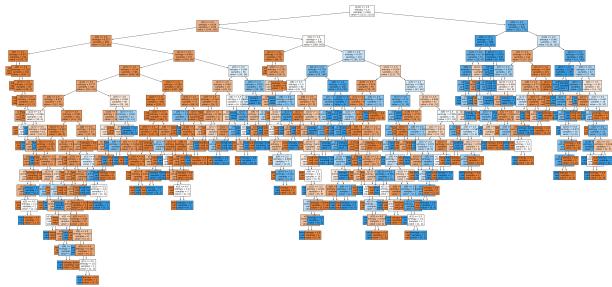
Hình 69. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật SMOTE và Randomized Search.



Hình 70. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật US theo kiểu Default value.



Hình 73. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật US theo kiểu Default value.



Hình 71. Biểu đồ biểu diễn cây quyết định khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật US theo kiểu Default value.

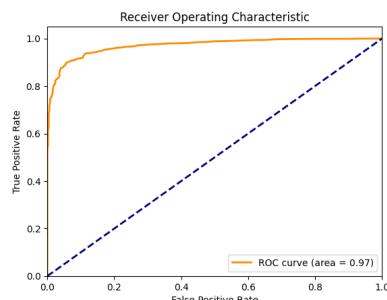
tham gia tín dụng như Hình 84.

c) Thuật toán XGBoost

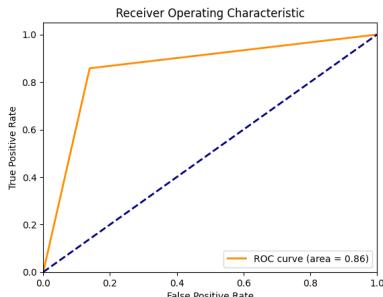
Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 85.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 86.

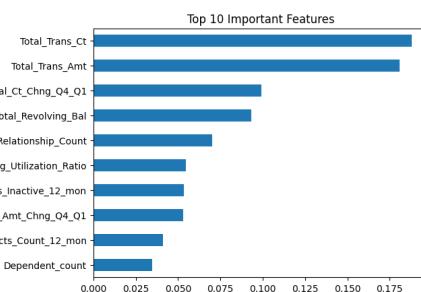
Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc



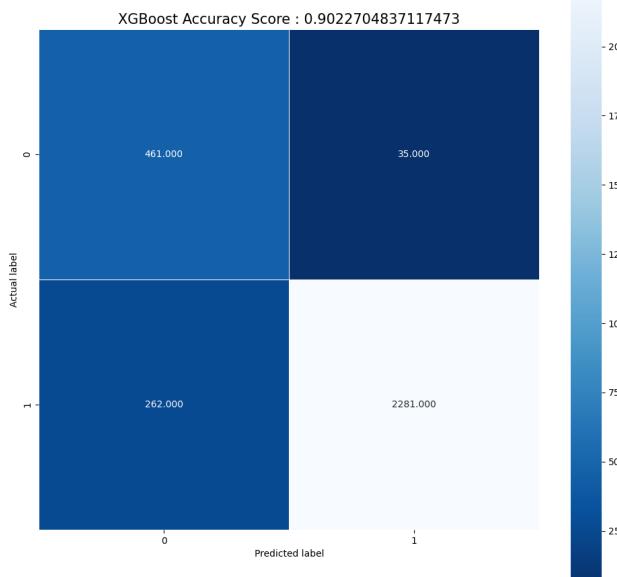
Hình 74. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật US theo kiểu Default value.



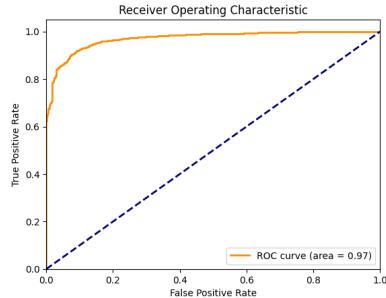
Hình 72. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật US theo kiểu Default value.



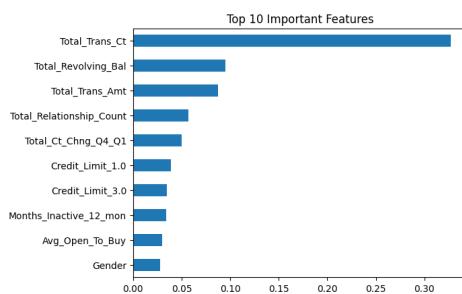
Hình 75. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật US theo kiểu Default value.



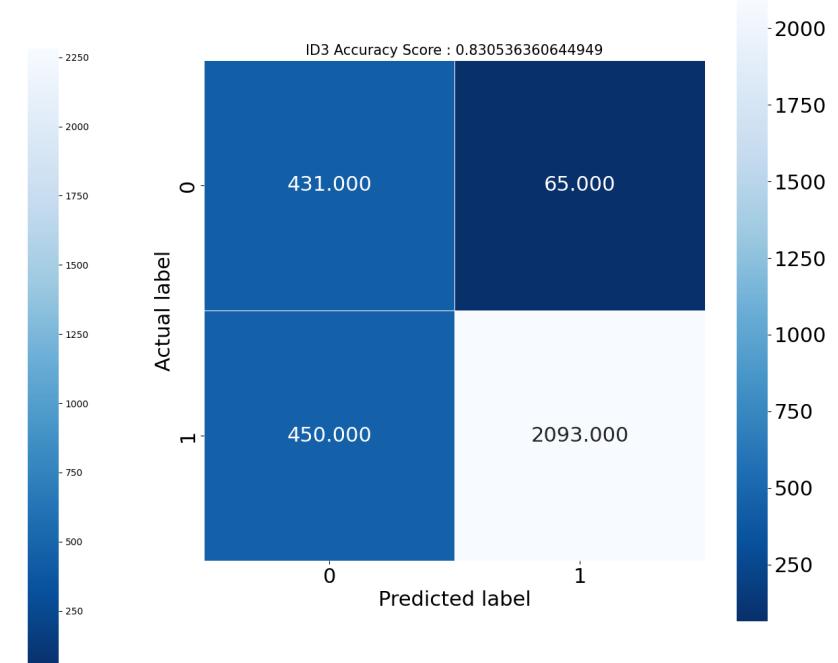
Hình 76. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật US theo kiểu Default value.



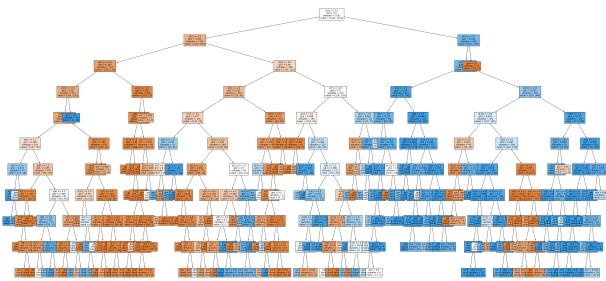
Hình 77. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật US theo kiểu Default value.



Hình 78. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật US theo kiểu Default value.



Hình 79. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật US và Randomized Search.



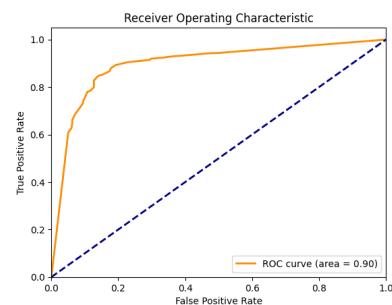
Hình 80. Biểu đồ biểu diễn cây quyết định khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật US và Randomized Search.

tham gia tín dụng như Hình 87.

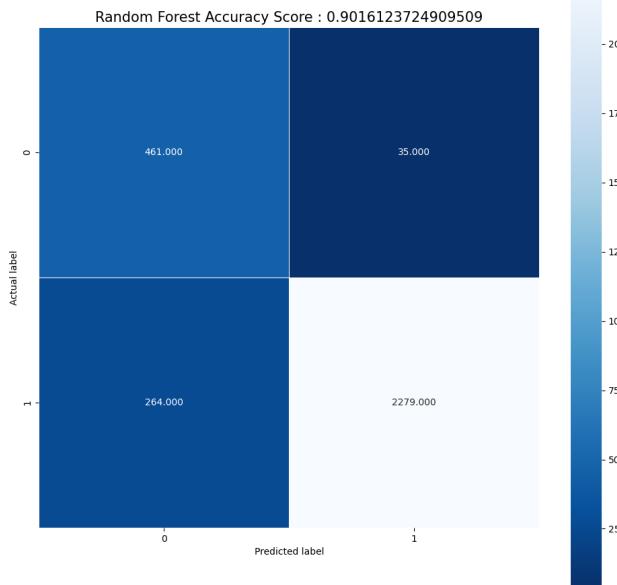
7) Thực hiện chia tập dữ liệu 8-2 và dùng kỹ thuật Oversampling để cân bằng dữ liệu theo kiểu Default value và tính toán các thông số đánh giá:

a) Thuật toán ID3

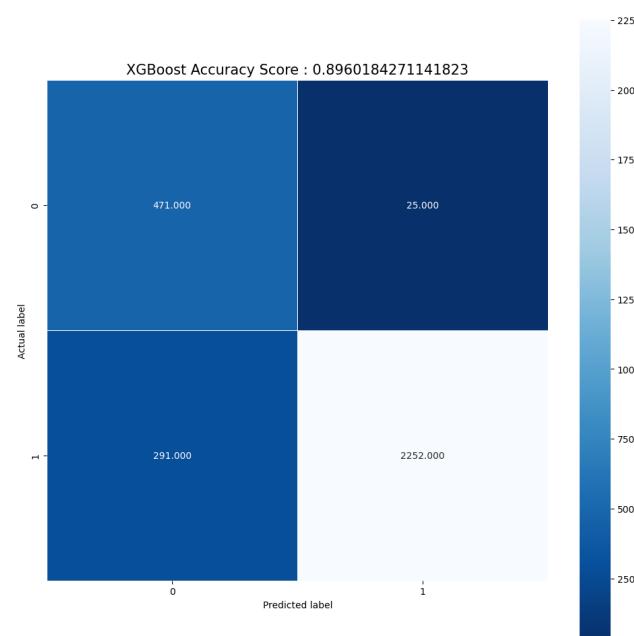
Thông số đánh giá Accuracy theo kiểu Default value được



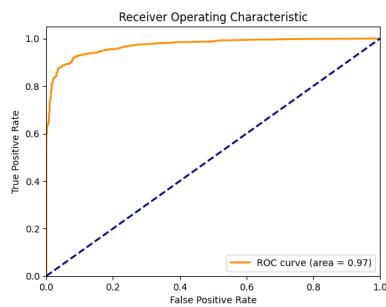
Hình 81. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán ID3 với kỹ thuật US và Randomized Search.



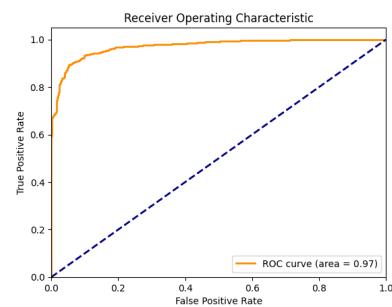
Hình 82. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật US và Randomized Search.



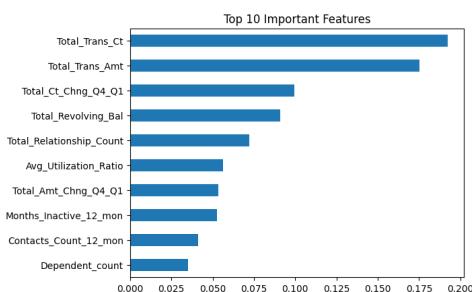
Hình 85. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật US và Randomized Search.



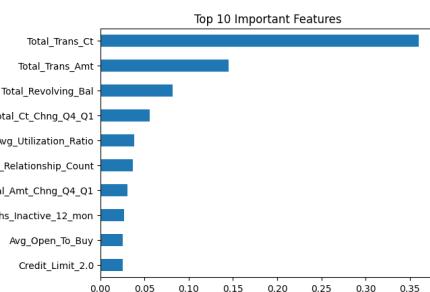
Hình 83. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật US và Randomized Search.



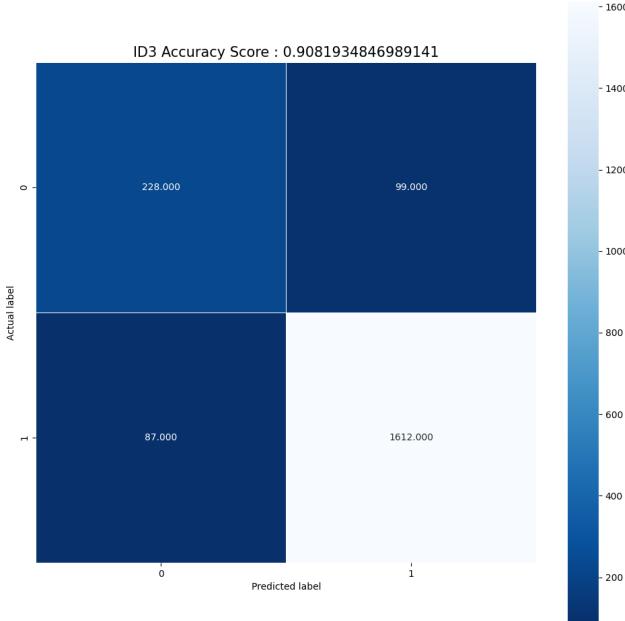
Hình 86. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật US và Randomized Search.



Hình 84. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán RF với kỹ thuật US và Randomized Search.



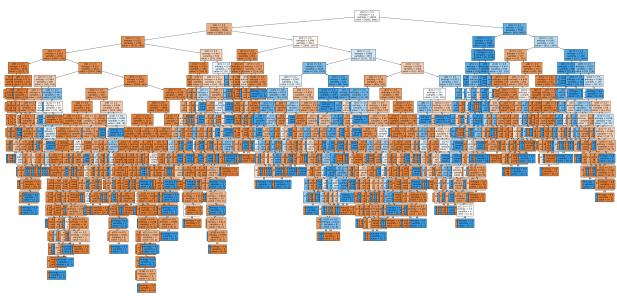
Hình 87. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 7-3 dùng thuật toán XGBoost với kỹ thuật US và Randomized Search.



Hình 88. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật OS theo kiểu Default value.

biểu diễn như Hình 88.

Biểu đồ biểu diễn cây quyết định như Hình 89.



Hình 89. Biểu đồ biểu diễn cây quyết định khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật OS theo kiểu Default value.

Đường cong ROC theo kiểu Default value được biểu diễn như Hình 90.

b) Thuật toán RF

Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 91.

Đường cong ROC theo kiểu Default value được biểu diễn như Hình 92.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 93.

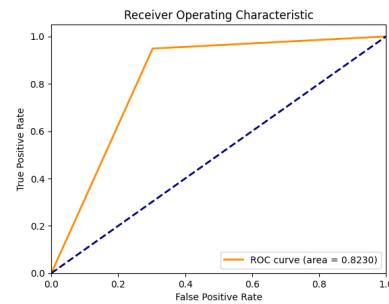
c) Thuật toán XGBoost

Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 94.

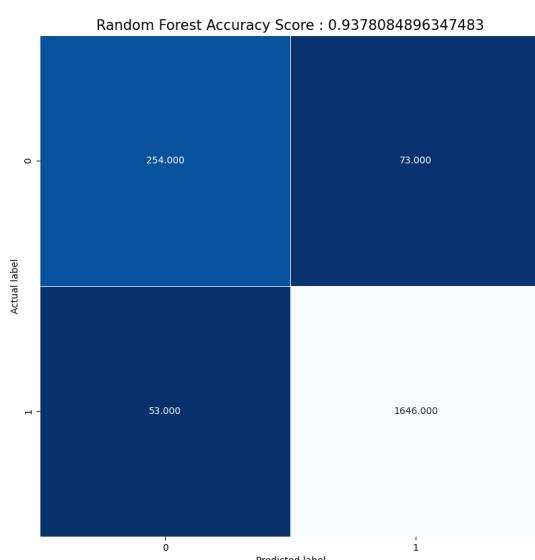
Đường cong ROC theo kiểu Default value được biểu diễn như Hình 95.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 96.

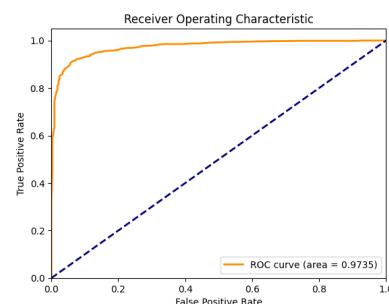
8) Thực hiện chia tập dữ liệu 8-2 và dùng kỹ thuật Oversampling để cân bằng dữ liệu và kết hợp Randomized Search để tính toán các thông số đánh giá:



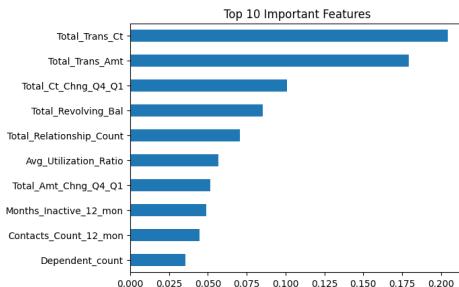
Hình 90. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật OS theo kiểu Default value.



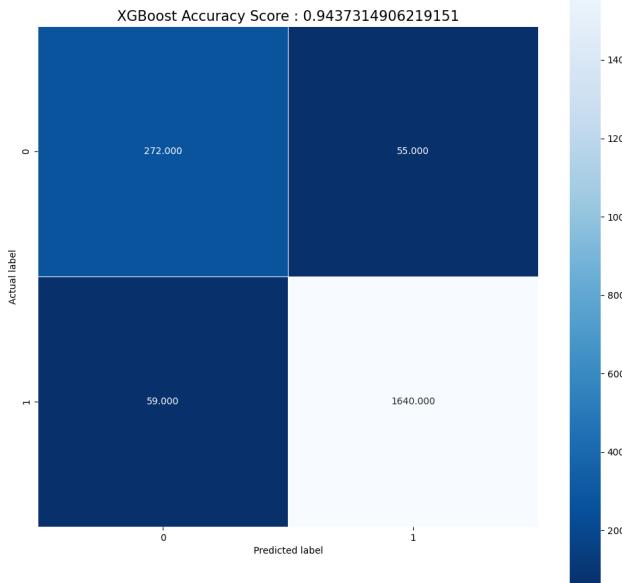
Hình 91. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật OS theo kiểu Default value.



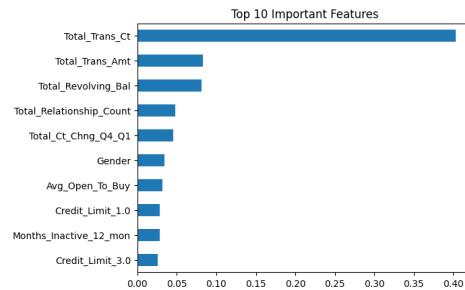
Hình 92. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật OS theo kiểu Default value.



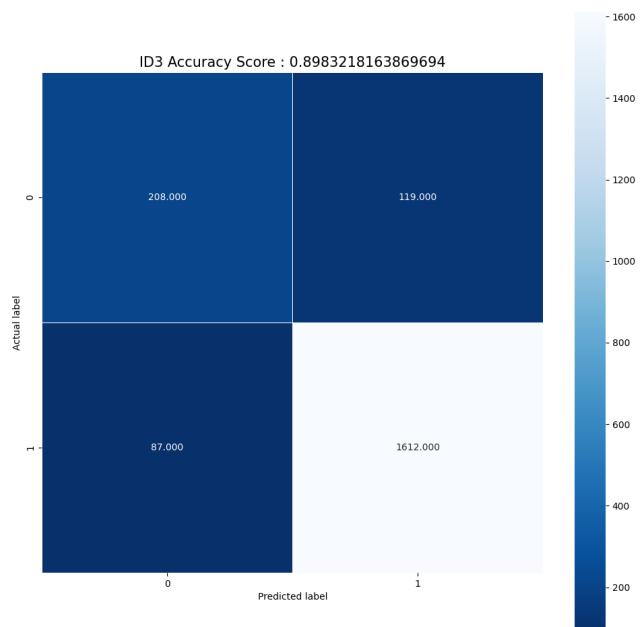
Hình 93. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật OS theo kiểu Default value.



Hình 94. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật OS theo kiểu Default value.



Hình 95. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật OS theo kiểu Default value.

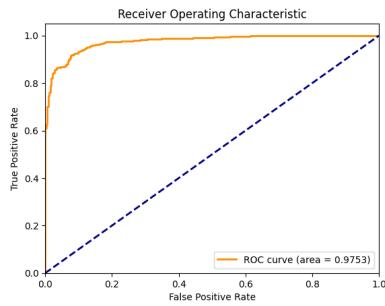


Hình 96. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật OS và Randomized Search.

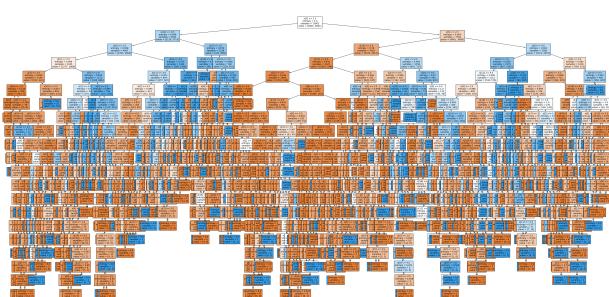
a) Thuật toán ID3

Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 97.

Biểu đồ biểu diễn cây quyết định như Hình 98



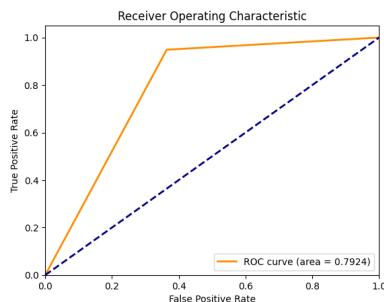
Hình 97. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật OS và Randomized Search.



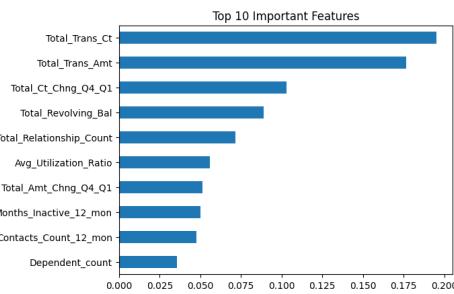
Hình 98. Biểu đồ biểu diễn cây quyết định khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật OS và Randomized Search.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 99.

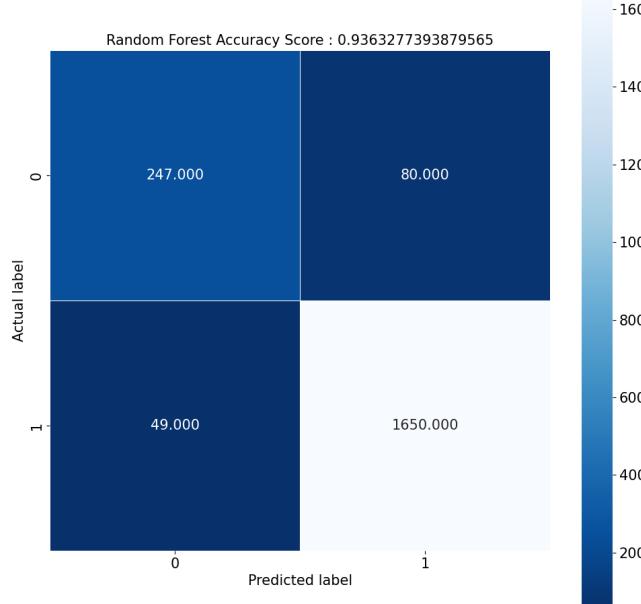
b) Thuật toán RF



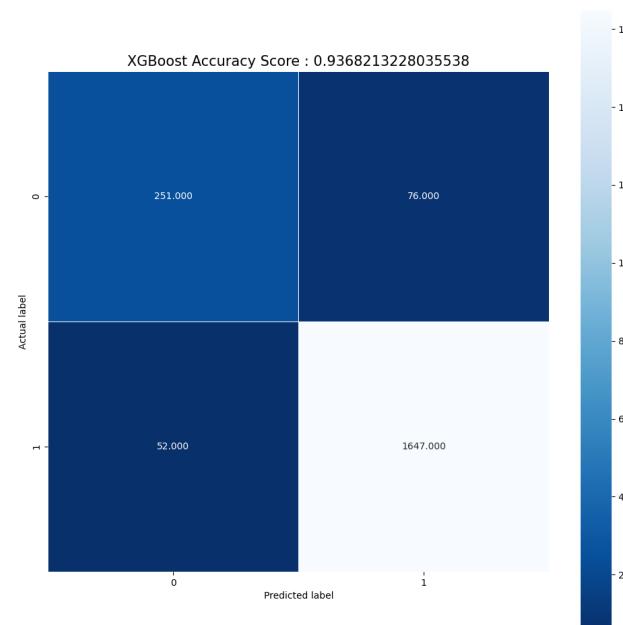
Hình 99. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật OS và Randomized Search.



Hình 102. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật OS và Randomized Search.



Hình 100. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật OS và Randomized Search.



Hình 103. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật OS và Randomized Search.

Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 100.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 101.

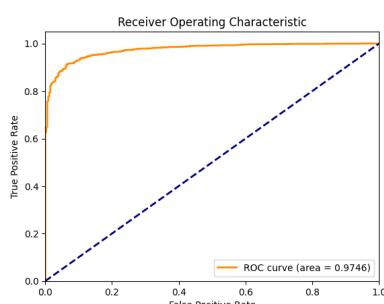
Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 102.

c) Thuật toán XGBoost

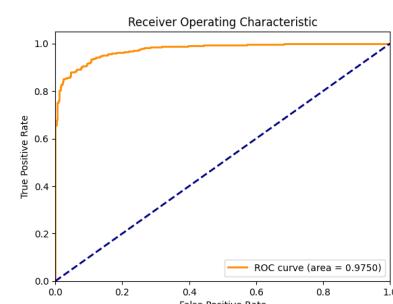
Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 103.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 104.

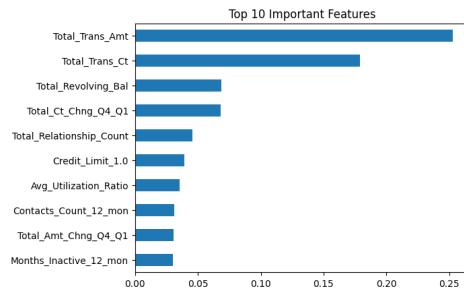
Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc



Hình 101. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật OS và Randomized Search.



Hình 104. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật OS và Randomized Search.



Hình 105. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật OS và Randomized Search.



Hình 106. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật SMOTE theo kiểu Default value.

tham gia tín dung như Hình 105.

9) Thực hiện chia tập dữ liệu 8-2 và dùng kỹ thuật SMOTE để cân bằng dữ liệu theo kiểu Default value và tính toán các thông số đánh giá:

a) Thuật toán ID3

Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 106.

Biểu đồ biểu diễn cây quyết định như Hình 107.

Đường cong ROC theo kiểu Default value được biểu diễn như Hình 108.

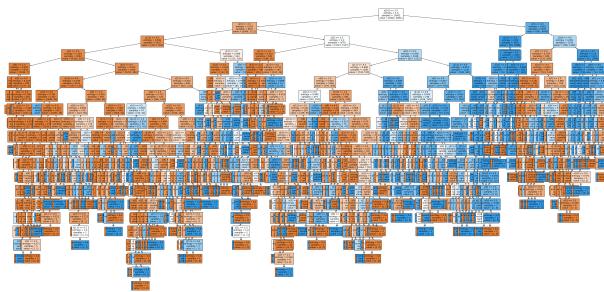
b) Thuật toán RF

Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 109.

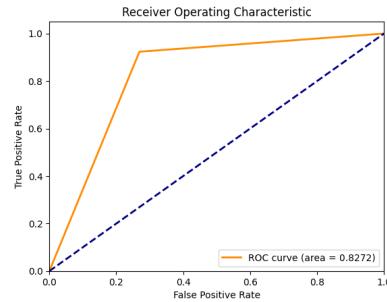
Đường cong ROC theo kiểu Default value được biểu diễn như Hình 110.

Biểu đồ biểu diễn Top

Bảng số liệu dịch Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 111.



Hình 107. Biểu đồ biểu diễn cây quyết định khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật SMOTE theo kiểu Default value.

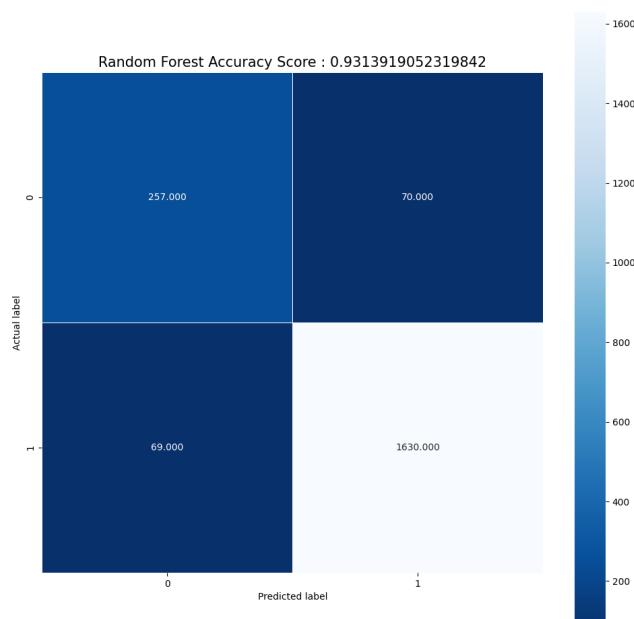


Hình 108. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật SMOTE theo kiểu Default value.

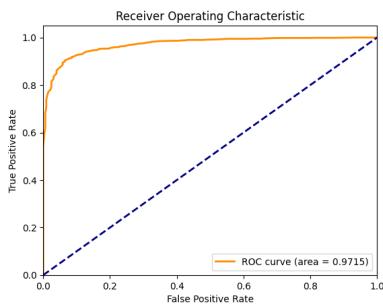
Đường cong ROC theo kiểu Default value được biểu diễn như Hình 113.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 114.

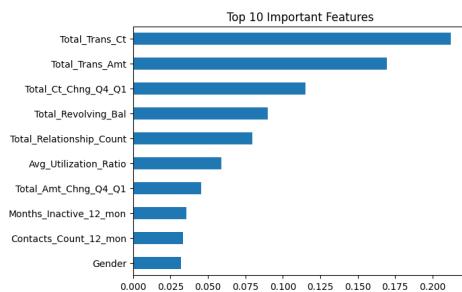
10) Thực hiện chia tập dữ liệu 8-2 và dùng kỹ thuật SMOTE để cân bằng dữ liệu và kết hợp Randomized Search để tính toán các thông số đánh giá:



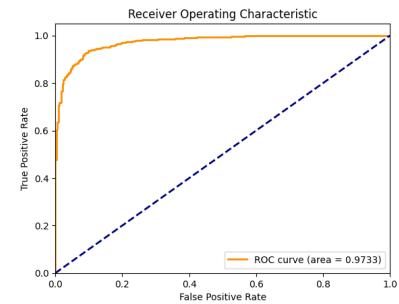
Hình 109. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật SMOTE theo kiểu Default value.



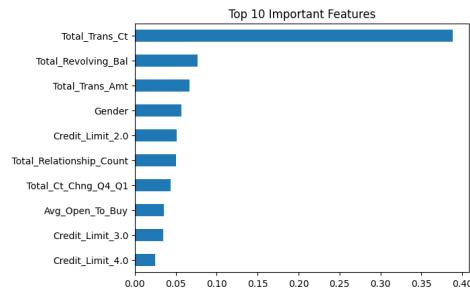
Hình 110. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật SMOTE theo kiểu Default value.



Hình 111. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật SMOTE theo kiểu Default value.



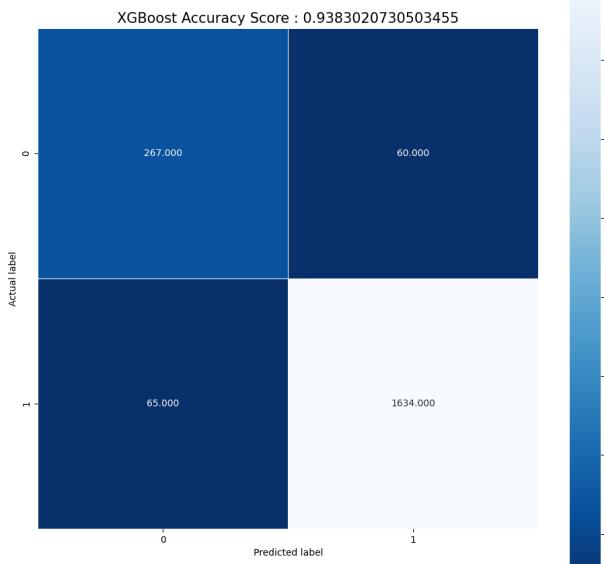
Hình 113. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật SMOTE theo kiểu Default value.



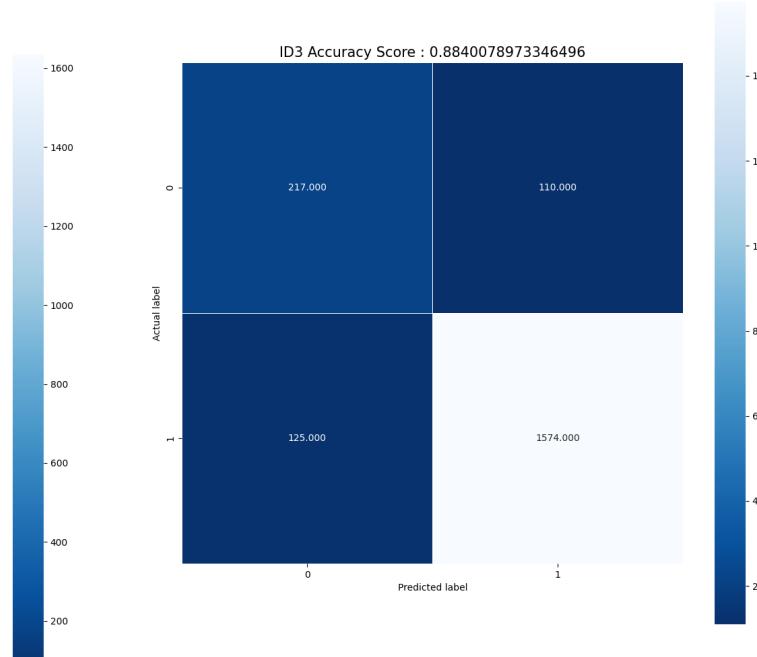
Hình 114. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2x dùng thuật toán XGBoost với kỹ thuật SMOTE theo kiểu Default value.

a) Thuật toán ID3

Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 115.



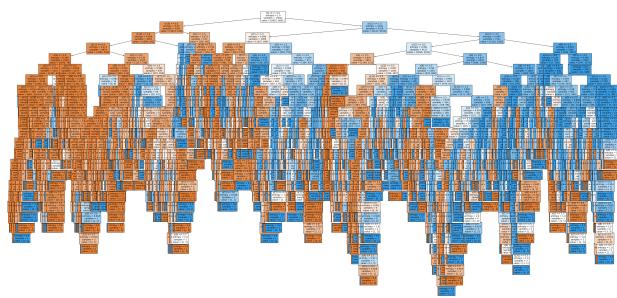
Hình 112. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật SMOTE theo kiểu Default value.



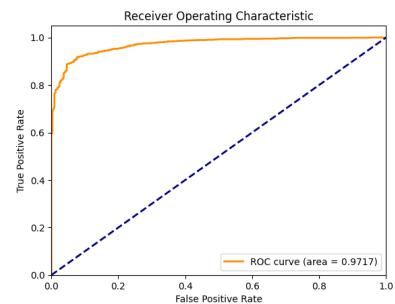
Hình 115. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật SMOTE và Randomized Search.

Biểu đồ biểu diễn cây quyết định như Hình 116.

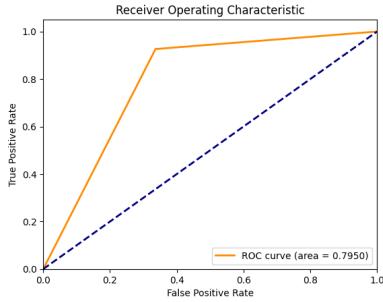
Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 117.



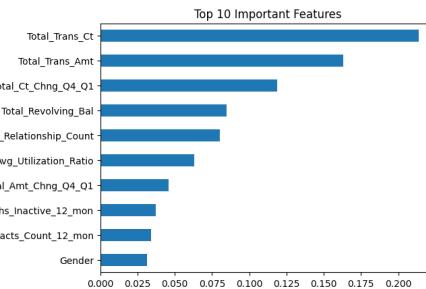
Hình 116. Biểu đồ biểu diễn cây quyết định khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật SMOTE và Randomized Search.



Hình 119. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật SMOTE và Randomized Search.



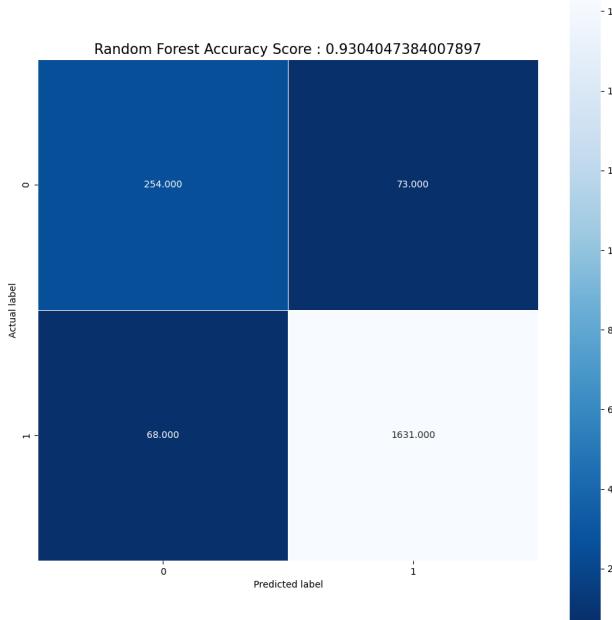
Hình 117. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật SMOTE và Randomized Search.



Hình 120. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật SMOTE và Randomized Search.

b) Thuật toán RF

Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 118.



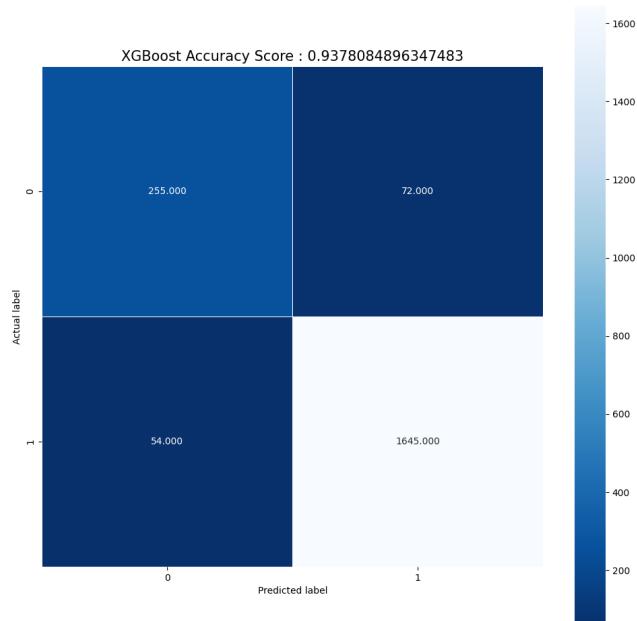
Hình 118. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật SMOTE và Randomized Search.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 119.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 120.

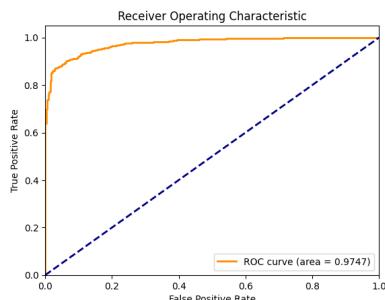
c) Thuật toán XGBoost

Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 121.



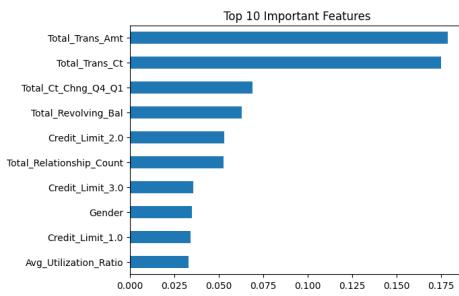
Hình 121. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật SMOTE và Randomized Search.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 122.



Hình 122. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật SMOTE và Randomized Search.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 123.



Hình 123. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật SMOTE và Randomized Search.

11) Thực hiện chia tập dữ liệu 8-2 và dùng kỹ thuật Undersampling để cân bằng dữ liệu theo kiểu Default value và tính toán các thông số đánh giá:

a) Thuật toán ID3

Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 124.

Biểu đồ biểu diễn cây quyết định như Hình 125.

Đường cong ROC theo kiểu Default value được biểu diễn như Hình 126.

b) Thuật toán RF

Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 127.

Đường cong ROC theo kiểu Default value được biểu diễn như Hình 128.

Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 129.

c) Thuật toán XGBoost

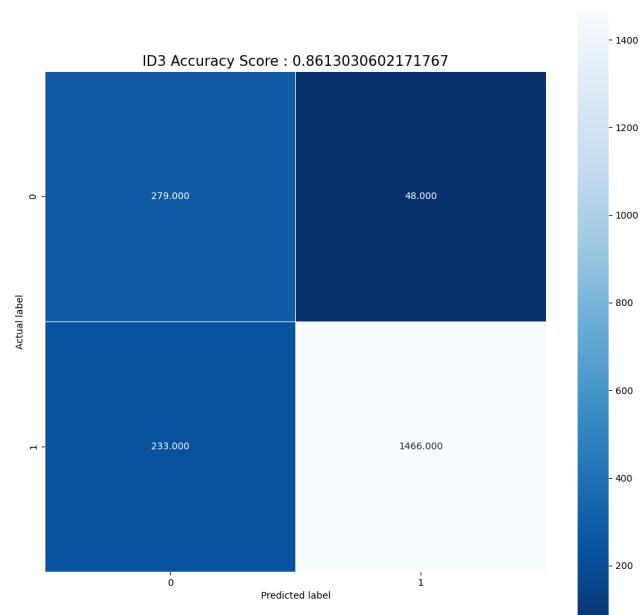
Thông số đánh giá Accuracy theo kiểu Default value được biểu diễn như Hình 130.

Đường cong ROC theo kiểu Default value được biểu diễn như Hình 131.

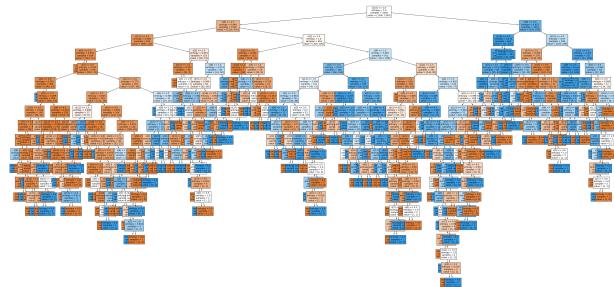
Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 132.

12) Thực hiện chia tập dữ liệu 8-2 và dùng kỹ thuật US để cân bằng dữ liệu và kết hợp Randomized Search để tính toán các thông số đánh giá:

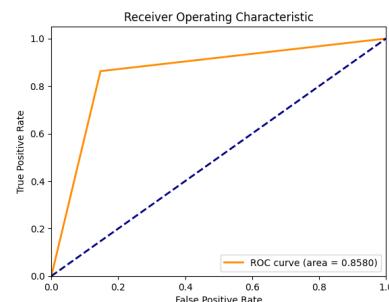
a) Thuật toán ID3



Hình 124. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật US theo kiểu Default value.



Hình 125. Biểu đồ biểu diễn cây quyết định khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật US theo kiểu Default value.



Hình 126. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật US theo kiểu Default value.

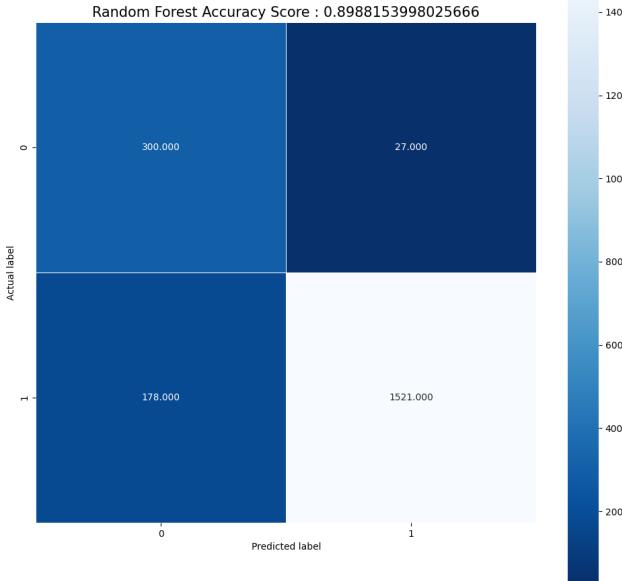
Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 133.

Biểu đồ biểu diễn cây quyết định như Hình 134

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 135.

b) Thuật toán RF

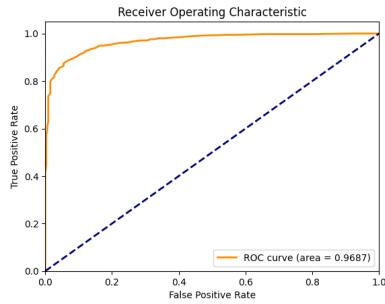
Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 136.



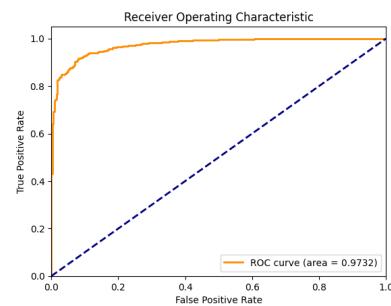
Hình 127. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật US theo kiểu Default value.



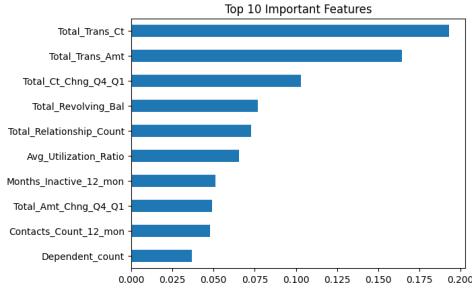
Hình 130. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật US theo kiểu Default value.



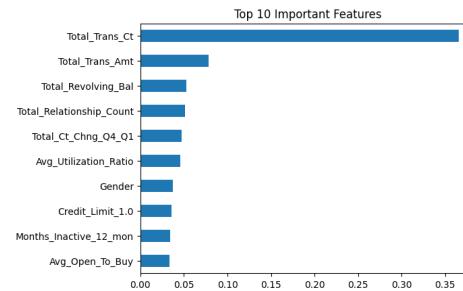
Hình 128. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật US theo kiểu Default value.



Hình 131. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật US theo kiểu Default value.



Hình 129. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi dùng chia tập dữ liệu 8-2 thuật toán RF với kỹ thuật US theo kiểu Default value.



Hình 132. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật US theo kiểu Default value.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 137.

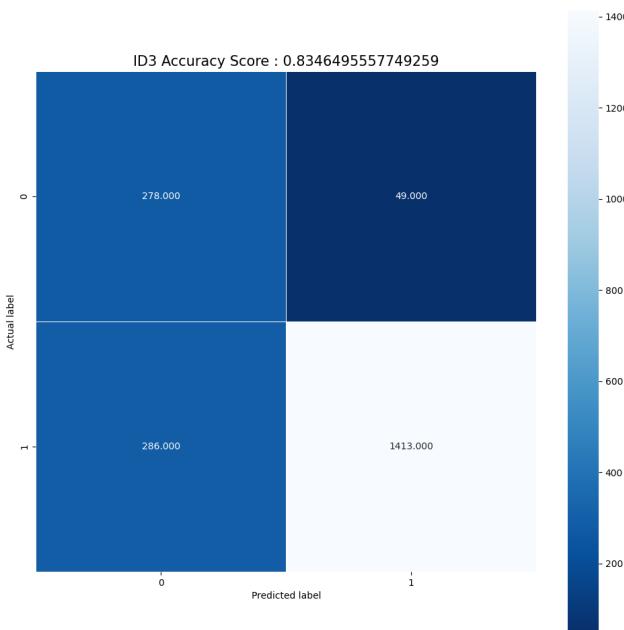
Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 138.

c) Thuật toán XGBoost

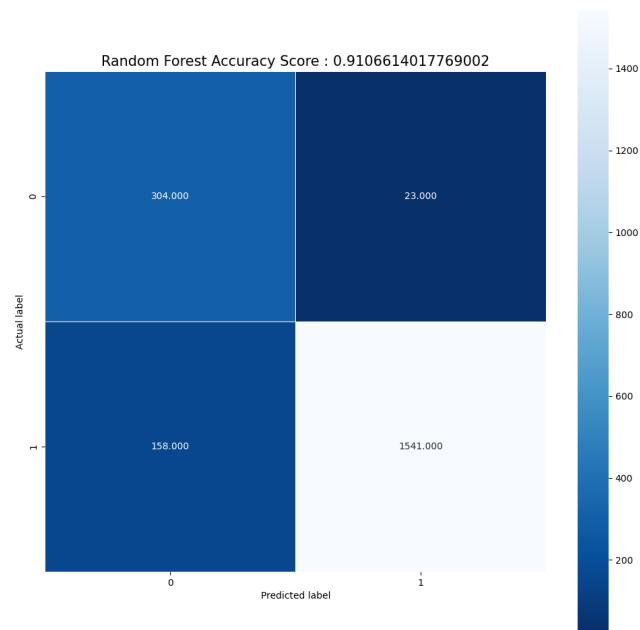
Thông số đánh giá Accuracy khi kết hợp với Randomized Search được biểu diễn như Hình 139.

Đường cong ROC khi kết hợp với Randomized Search được biểu diễn như Hình 140.

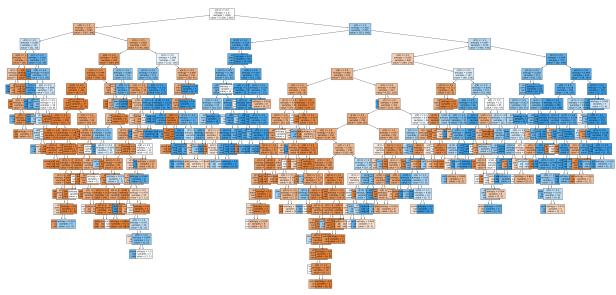
Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng như Hình 141.



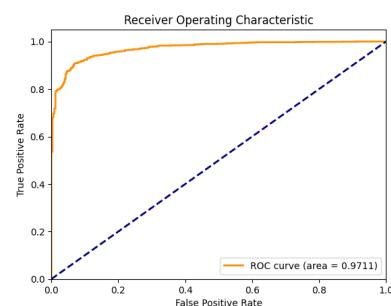
Hình 133. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật US và Randomized Search.



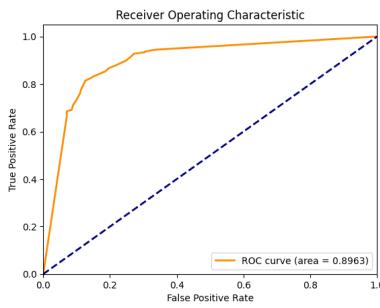
Hình 136. Biểu đồ biểu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật US và Randomized Search.



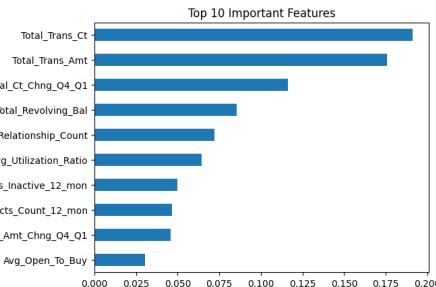
Hình 134. Biểu đồ biểu diễn cây quyết định khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật US và Randomized Search.



Hình 137. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật US và Randomized Search.



Hình 135. Biểu đồ biểu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán ID3 với kỹ thuật US và Randomized Search.



Hình 138. Biểu đồ biểu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2 dùng thuật toán RF với kỹ thuật US và Randomized Search.

C. So sánh thuật toán

a) Tham số mặc định

- Tỉ lệ dữ liệu 7-3:

Dựa trên kết quả của độ đo AUC có thể thấy với kiểu Default value (tức là không gắn các tham số mà để mặc định) thuật toán XGBoost và Random Forest đạt kết quả tốt nhất so với thuật toán ID3. Tuy nhiên, xét về chỉ số precision thì XGBoost lại có tỉ lệ cao hơn so với Random Forest. Vậy đối với tỉ lệ

dữ liệu 7-3 và đặt tham số mặc định thì XGBoost cho ra kết quả tốt nhất.

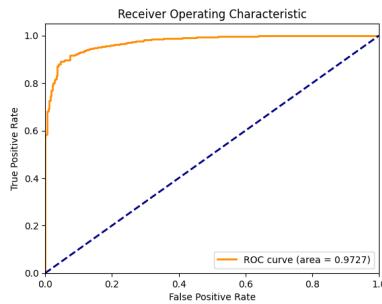
- Tỉ lệ dữ liệu 8-2:

Đối với tỉ lệ dữ liệu 8-2, dựa trên độ đo AUC và precision thì thuật toán XGBoost cho ra kết quả tốt nhất so với 2 thuật toán còn lại.

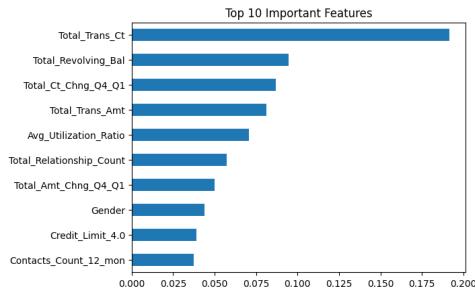
b) Sử dụng Randomized Search:



Hình 139. Biểu đồ biếu diễn giá trị thông số đánh giá Accuracy khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật US và Randomized Search.



Hình 140. Biểu đồ biếu diễn giá trị thông số đánh giá ROC khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật US và Randomized Search.



Hình 141. Biểu đồ biếu diễn Top 10 thuộc tính ảnh hưởng đến việc tham gia tín dụng khi chia tập dữ liệu 8-2 dùng thuật toán XGBoost với kỹ thuật US và Randomized Search.

- Tỉ lệ 7-3:

Khi sử dụng Randomized Search để tìm ra các tham số cho thuật toán thì dựa trên 2 chỉ số AUC và precision có thể thấy XGBoost có tỉ lệ chính xác cao hơn 2 thuật toán còn lại.

- Tỉ lệ 8-2:

Dựa trên 2 chỉ số AUC và precision thì thuật toán XGBoost cho ra kết quả tốt nhất.

c) Về các kĩ thuật xử lí mất cân bằng:

Đối với thuật toán ID3 thì kĩ thuật Undersampling cho ra độ chính xác tốt hơn so với 2 kĩ thuật SMOTE và Oversampling khoảng 1%.

Đối với thuật toán Random Forest và XGBoost, nhìn chung, kĩ thuật Oversampling cho ra kết quả tốt hơn 2 kĩ thuật SMOTE và Undersampling khoảng 1%.

d) Tổng quát:

Nhìn chung, đối với bài toán dự báo khách hàng ngừng tham gia tín dụng thì thuật toán XGBoost cho ra kết quả tốt nhất trong đa số các trường hợp. Ngược lại, ID3 lại cho ra kết quả không tốt bằng 2 thuật toán còn lại. Điều này cũng tương đối dễ hiểu khi ID3 là thuật toán nền tảng của các thuật toán dựa trên cây quyết định và XGBoost sinh ra đã khắc phục được phần lớn các khuyết điểm của thuật toán cây quyết định đời đầu.

VI. TỔNG KẾT VÀ HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI

Từ kết quả nghiên cứu, có thể thấy rằng việc khách hàng thường xuyên sử dụng thẻ tín dụng thì họ sẽ ít có khả năng sẽ ngừng sử dụng, những khách hàng có lịch sử giao dịch thường xuyên và khối lượng giao dịch lớn sẽ có tiềm năng hơn trong việc tham gia và tiếp tục duy trì các dịch vụ tín dụng, kết quả này tương đồng với nghiên cứu của X.Miao và đồng sự[4]. Với những khách hàng có tần suất giao dịch giảm dần có thể là những dấu hiệu đáng chú ý để các nhà hoạch định đưa ra những chính sách phù hợp để giữ chân khách hàng. Vì vậy những người quản trị ngân hàng có thể điều chỉnh các chương trình sử dụng thẻ tín dụng dựa trên tần suất khách hàng sử dụng thẻ để từ đó chống nguy cơ khách hàng đóng thẻ tín dụng và ngừng sử dụng dịch vụ.

Việc sử dụng các mô hình học máy dựa trên thuật toán cây quyết định như ID3, Random Forest, XGBoost có kết quả dự đoán ở mức tốt. Tuy nhiên việc sử dụng các thuật toán đơn thuần sẽ chưa thực sự đạt hiệu quả nếu dữ liệu đầu vào mất cân bằng nên việc áp dụng các cơ chế xử lí mất cân bằng dữ liệu như SMOTE, Undersampling hay Oversampling sẽ đem lại hiệu quả tốt hơn.

Bài nghiên cứu đã đáp ứng được tiêu chí ban đầu là có thể xây dựng và so sánh được hiệu quả của các thuật toán trong việc dự báo phân loại việc khách hàng ngừng tham gia tín dụng. Tuy nhiên, việc chỉ thực nghiệm trên một bộ dữ liệu sẽ đem lại kết quả chủ quan và chưa thực tế. Trong thực tế, để phát triển nghiên cứu lên mức ứng dụng vào thực tiễn cần phải thực nghiệm thêm trên những bộ dữ liệu thực tế từ nhiều tổ chức tài chính/ngân hàng cũng như thực nghiệm trên nhiều mô hình thuật toán hơn. Vì thế trong tương lai, nhóm sẽ cố gắng phát triển bài nghiên cứu ở mức độ doanh nghiệp để có thể áp dụng vào thực tế lĩnh vực ngân hàng.

VII. LỜI CẢM ƠN

Cuối cùng, nhóm tác giả xin gửi lời cảm ơn đến giảng viên hướng dẫn Hà Lê Hoài Trung đã luôn hỗ trợ và góp ý để nhóm có thể hoàn thành bài nghiên cứu một cách tốt nhất.

Thang do		Accuracy		F1		Recall		Precision		AUC	
Tí lệ dữ liệu	7-3	8-2	7-3	8-2	7-3	8-2	7-3	8-2	7-3	8-2	
ID3	SMOTE	0.8960	0.8923	0.70	0.69	0.74	0.73	0.66	0.65	0.8340	0.8272
	US	0.8585	0.8613	0.67	0.67	0.86	0.85	0.54	0.54	0.8595	0.8580
	OS	0.9118	0.9082	0.72	0.71	0.71	0.70	0.74	0.72	0.8313	0.8230
RF	SMOTE	0.9335	0.9314	0.80	0.79	0.81	0.79	0.79	0.79	0.9703	0.9715
	US	0.9052	0.8988	0.76	0.75	0.94	0.92	0.64	0.63	0.9707	0.9687
	OS	0.9385	0.9378	0.80	0.80	0.77	0.78	0.84	0.83	0.9735	0.9735
XGB	SMOTE	0.9395	0.9383	0.81	0.83	0.81	0.83	0.82	0.82	0.9739	0.9733
	US	0.9023	0.9121	0.76	0.81	0.93	0.82	0.64	0.80	0.9708	0.9732
	OS	0.9388	0.9437	0.82	0.77	0.84	0.92	0.80	0.66	0.9754	0.9753

Bảng I

BẢNG SO SÁNH HIỆU QUẢ THUẬT TOÁN THEO KIỂU DEFAULT VALUE

Thang do		Accuracy		F1		Recall		Precision		AUC	
Tí lệ dữ liệu	7-3	8-2	7-3	8-2	7-3	8-2	7-3	8-2	7-3	8-2	
ID3	SMOTE	0.8809	0.8840	0.65	0.65	0.68	0.66	0.62	0.63	0.7998	0.7950
	US	0.8305	0.8346	0.63	0.62	0.87	0.85	0.49	0.49	0.9045	0.8963
	OS	0.8960	0.8983	0.67	0.67	0.66	0.64	0.69	0.71	0.7960	0.7924
RF	SMOTE	0.9358	0.9304	0.80	0.78	0.80	0.78	0.81	0.79	0.9720	0.9717
	US	0.9065	0.9107	0.77	0.77	0.94	0.93	0.65	0.66	0.9727	0.9711
	OS	0.9395	0.9363	0.80	0.79	0.76	0.76	0.85	0.83	0.9748	0.9747
XGB	SMOTE	0.9368	0.9378	0.80	0.80	0.78	0.78	0.82	0.83	0.9750	0.9747
	US	0.9006	0.9160	0.75	0.78	0.94	0.94	0.63	0.68	0.9696	0.9727
	OS	0.9401	0.9368	0.81	0.80	0.79	0.77	0.84	0.83	0.9771	0.9750

Bảng II

BẢNG SO SÁNH HIỆU QUẢ THUẬT TOÁN VỚI RANDOMIZED SEARCH

TÀI LIỆU

- [1] M. J. Bernthal, D. Crockett, and R. L. Rose, “Credit cards as lifestyle facilitators,” *Journal of Consumer Research*, vol. 32, no. 1, pp. 130–145, Jun. 2005. doi: 10.1086/429605. [Online]. Available: <https://doi.org/10.1086/429605>.
- [2] X. Zhu, W. Ren, Q. Chen, and R. Evans, “How does internet usage affect the credit consumption among chinese college students? a mediation model of social comparison and materialism,” *Internet Research*, vol. 31, no. 3, pp. 1083–1101, Nov. 2020. doi: 10.1108/intr-08-2019-0357. [Online]. Available: <https://doi.org/10.1108/intr-08-2019-0357>.
- [3] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, “Computer assisted customer churn management: State-of-the-art and future trends,” *Computers & Operations Research*, vol. 34, no. 10, pp. 2902–2917, Oct. 2007. doi: 10.1016/j.cor.2005.11.007. [Online]. Available: <https://doi.org/10.1016/j.cor.2005.11.007>.
- [4] X. Miao and H. Wang, “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)*, Atlantis Press, 2022, pp. 649–656, ISBN: 978-94-6239-554-1. doi: 10.2991/aebmr.k.220307.104. [Online]. Available: <https://doi.org/10.2991/aebmr.k.220307.104>.
- [5] M. Kaur, K. Singh, and N. Sharma, “Data mining as a tool to predict the churn behaviour among indian bank customers,” 2013.
- [6] S. Agrawal, A. Das, A. Gaikwad, and S. Dhage, “Customer churn prediction modelling based on behavioural patterns analysis using deep learning,” in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, 2018, pp. 1–6. doi: 10.1109/ICSCCEE.2018.8538420.
- [7] A. S. Choudhari and M. Potey, “Predictive to prescriptive analysis for customer churn in telecom industry using hybrid data mining techniques,” in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBECA)*, IEEE, Aug. 2018. doi: 10.1109/iccubea.2018.8697532. [Online]. Available: <https://doi.org/10.1109/iccubea.2018.8697532>.
- [8] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, “A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector,” *IEEE Access*, vol. 7, pp. 60134–60149, 2019. doi: 10.1109/ACCESS.2019.2914999.
- [9] V. P. Malikireddy and M. Kasa, “Customer churns prediction model based on machine learning techniques: A systematic review,” *Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021)*, 2021.
- [10] B. Raja and P. Jeyakumar, “An effective classifier for predicting churn in telecommunication,” *Journal of Advanced Research in Dynamical and Control Systems*, vol. 11, pp. 221–229, Jun. 2019.
- [11] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, “Customer churn prediction system: A machine learning approach,” *Computing*, vol. 104, no. 2, pp. 271–294, Feb. 2021. doi: 10.1007/s00607-021-00908-y. [Online]. Available: <https://doi.org/10.1007/s00607-021-00908-y>.
- [12] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine learning with oversampling and undersampling techniques: Overview study and experimental results,”

- in 2020 11th International Conference on Information and Communication Systems (ICICS), IEEE, Apr. 2020. DOI: 10 . 1109 / icics49469 . 2020 . 239556. [Online]. Available: <https://doi.org/10.1109/icics49469.2020.239556>.
- [13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," 2011. DOI: 10.48550/ARXIV.1106.1813. [Online]. Available: <https://arxiv.org/abs/1106.1813>.
- [14] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 10, pp. 281–305, 2012. [Online]. Available: <http://jmlr.org/papers/v13/bergstra12a.html>.
- [15] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986. DOI: 10.1007/bf00116251. [Online]. Available: <https://doi.org/10.1007/bf00116251>.
- [16] A. Rajeshkanna and K. Arunesh, "Id3 decision tree classification: An algorithmic perspective based on error rate," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 787–790. DOI: 10.1109/ICESC48915.2020.9155578.
- [17] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001. DOI: 10 . 1023 / A : 1010950718922.
- [18] D. Steinberg, "Cart: Classification and regression trees," 2009.
- [19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, San Francisco, California, USA: ACM, 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. [Online]. Available: <http://doi.acm.org/10.1145/2939672.2939785>.