DỰ ĐOÁN TỶ LỆ RỜI BỎ KHÁCH HÀNG CỦA NGÂN HÀNG BẰNG MÁY HỌC
MACHINE LEARNING FOR BANKING CUSTOMER CHURN PREDICTION

Trần Thành Công
Khoa Công nghệ thông tin, trường Đại học Kinh tế tài chính Hồ Chí Minh

Tóm tắt*: Khách hàng rời bỏ hiện đang trở thành một vấn đề nghiêm trọng trong lĩnh vực ngân hàng. Cần tìm giải pháp dự đoán tỷ lệ khách hàng rời bỏ ngân hàng; tuy nhiên, bộ dữ liệu cho dự đoán khách hàng rời bỏ ngân hàng là không cân bằng. Trong bài báo này, Rừng ngẫu nhiên (RF) dựa trên kỹ thuật lấy mẫu lại phổ biến có tên là SMOTE được sử dụng để tạo ra mô hình dự đoán rời bỏ khách hàng của ngân hàng. Các chỉ số, bao gồm độ chính xác, thu hồi, độ đặc hiệu, điểm F1, hệ số tương quan Mathews và ROC-AUC được sử dụng để đánh giá toàn diện mô hình dự đoán. Qua kết quả thực nghiệm, các giá trị Accuracy và ROC-AUC của mô hình RF dựa trên SMOTE kết quả khả quan.*

Từ khóa*: Khách hàng rời bỏ, Ngành ngân hàng; Rừng ngẫu nhiên, SMOTE.*

Abstract*: Customer Churn is now becoming a significant problem in banking sector. It is necessary to seek solutions to predict the rate of customer churn in banks; however, dataset for customer churn in banks prediction is imbalanced. In this paper, Random Forest (RF) based on two popular resampling techniques, named SMOTE is used to obtain banking customer churn prediction model. A wide range of metrics, including Accuracy, Recall, Specificity, F1 score, Mathews correlation coefficient and ROC-AUC, are used to comprehensively evaluation the prediction model. Through the experimental results, the values Accuracy and ROC-AUC of RF model based on SMOTE indicate positive results.*

Keywords*: Customer churn, Banking sector; Random Forest, SMOTE.*

1. Giới thiệu

Hiện nay, mục đích làm tăng sự hài lòng của khách hàng là một trong những mục đích quan trọng của các ngân hàng trên toàn thế giới. Ngày nay, khách hàng có xu hướng ứng dụng công nghệ mới vào nhiều mặt của cuộc sống, trong đó có dịch vụ ngân hàng. Điều này dẫn đến sự cạnh tranh gay gắt giữa các ngân hàng để giữ chân khách hàng. Vì vậy, nhiều ngân hàng trên thế giới cần phải tìm nhiều cách để hạn chế tỷ lệ rời bỏ khách hàng. Khách hàng rời bỏ được định nghĩa là việc khách hàng hiện đang sử dụng dịch vụ của ngân hàng này rời bỏ để chuyển sang sử dụng dịch vụ của ngân hàng đối thủ khác.

Ngày nay, vấn đề khách hàng rời bỏ ngân hàng ngày càng trở nên phổ biến. Nhiều nghiên cứu đã chỉ ra rằng việc loại bỏ tình trạng rời bỏ khách hàng có thể tiết kiệm một khoản tiền lớn vì việc có được khách hàng mới thường tốn gấp năm lần so với việc đáp ứng và giữ chân những khách hàng hiện có. Do đó, để tránh tình trạng khách hàng rời bỏ, các ngân hàng đã đầu tư thiết lập hệ thống quản lý quan hệ khách hàng để thu thập dữ liệu, phân tích hành vi của khách hàng và đề xuất các kỹ thuật giữ chân khách hàng [1].

Tuy nhiên, có một số thách thức trong việc xác định sự rời bỏ trong lĩnh vực ngân hàng. Thứ nhất, hiện có hàng triệu khách hàng đang sử dụng dịch vụ ngân hàng tại các ngân hàng lớn, đặc biệt là các ngân hàng quốc tế nên việc thu thập đầy đủ bộ dữ liệu mất nhiều thời gian và bộ dữ liệu thu thập không được tổng hợp tốt. Thứ hai, ngân hàng có thể không có khả năng thích ứng kịp thời với những thay đổi về nhu cầu của khách hàng. Thứ ba, việc xác định các mẫu khách hàng theo cách thủ công vẫn là một thách thức đối với nhân viên ngân hàng mặc dù các ngân hàng phân khúc khách hàng theo các nhà quản lý địa phương. Do đó, điều quan

trọng là phải chuyển từ các phương pháp tiếp cận truyền thống sang các phương pháp hiện đại, chẳng hạn như ứng dụng thuật toán học máy (ML) để phân tích hành vi của khách hàng và tìm ra giải pháp ngăn chặn sự rời bỏ của khách hàng.

Nghiên cứu này nghiên cứu hành vi của một bộ dữ liệu đại diện gồm 10.000 khách hàng thu được từ Kaggle để tạo ra một mô hình dự đoán rời bỏ thông qua thuật toán Rừng ngẫu nhiên (RF) dựa trên hai kỹ thuật lấy mẫu lại phổ biến, đó là SMOTE. Hiệu suất của mô hình dự đoán rời bỏ khách hàng của ngân hàng được đánh giá thông qua nhiều chỉ số khác nhau. Ngoài ra, các tính năng được sắp xếp từ quan trọng nhất đến ít quan trọng hơn cũng được xác định thông qua thuật toán RF.

Nghiên cứu này được tổ chức như sau: Phần 2 trình bày tổng quan tài liệu. Phần 3 mô tả phương pháp luận. Phần 4 trình bày kết quả của nghiên cứu này. Phần 5 tiết lộ kết luận và công việc trong tương lai.

2. Các công trình nghiên cứu liên quan

Một loạt các thuật toán ML, bao gồm Cây quyết định (DT), Rừng ngẫu nhiên (RF), K-Láng giềng gần nhất (KNN), Hồi quy logistic (LR) và Máy vectơ hỗ trợ (SVM) đã được nhiều nhà nghiên cứu đề xuất cho khách hàng rời đi. dự đoán trong nhiều thập kỷ qua.

Trong [2] , các kỹ thuật ML như KNN, SVM, DT và RF đã được sử dụng để dự đoán tỷ lệ khách hàng rời bỏ ngân hàng. Dựa trên tập dữ liệu của 10000 khách hàng ngân hàng thu được từ Kaggle, nghiên cứu này đã phân tích hành vi của khách hàng để khám phá khả năng rời bỏ. Để tăng độ chính xác của các mô hình ML, một số phương pháp lựa chọn tính năng đã được triển khai để xác định các tính năng có liên quan. Kết quả của mô hình RF dựa trên phương pháp oversampling tốt hơn các mô hình khác về chỉ số độ chính xác.

Trong [3], sáu cách tiếp cận khác nhau, bao gồm RF, SVM, Tăng cường ngẫu nhiên (SB), LR, Cây hồi quy (GIỎI), Đường hồi quy thích ứng đa biến (MARS) đã được sử dụng để dự đoán sự thay đổi của ngân hàng bán lẻ. Nghiên cứu này đã phát triển một khung phương pháp luận có khả năng dự đoán cả việc khách hàng nào ngừng sử dụng dịch vụ ngân hàng và thời điểm họ ngừng sử dụng dịch vụ ngân hàng trong khoảng thời gian sáu tháng trong tương lai. Nghiên cứu này chỉ ra rằng SB chứng minh tính hiệu quả của nó so với các phương pháp tiếp cận khác dựa trên bộ dữ liệu của hơn 130 000 khách hàng của một ngân hàng bán lẻ. Ngoài ra, nghiên cứu này cho thấy hai biến có tên là tổng giá trị sản phẩm ngân hàng nắm giữ trong những tháng gần đây và sự tồn tại của thẻ ghi nợ hoặc thẻ tín dụng ở một ngân hàng khác có tác động tích cực đến dự đoán rời bỏ trong khoảng thời gian 1-2 tháng. Nhưng hai biến số khác là số lượng giao dịch trong những tháng gần đây và sự tồn tại của một khoản vay thế chấp bên ngoài ngân hàng đóng một vai trò quan trọng trong dự đoán churn trong khoảng thời gian 3-4 tháng.

Trong [4], các cách tiếp cận khác nhau là LR, bayes ngây thơ, SVM, RF, DT và các kỹ thuật tăng cường và tập hợp đã được sử dụng để dự đoán tỷ lệ khách hàng rời bỏ dựa trên bộ dữ liệu viễn thông. Nghiên cứu này đề xuất một khuôn khổ bao gồm sáu giai đoạn i, e. tiền xử lý dữ liệu, phân tích tính năng, lựa chọn tính năng, quy trình phân tách, quy trình dự đoán và xác thực. Trong nghiên cứu này, trong số các phương pháp nói trên, Adaboost và XGboost Classifier có kết quả cao nhất về độ chính xác và điểm AUC.

Trong [5], bài báo đã phát triển một mô hình dự đoán rời bỏ bằng cách sử dụng ML để hỗ trợ các nhà khai thác viễn thông dự đoán những khách hàng có khả năng rời bỏ dịch vụ. Mô hình này được tạo ra dựa trên cách tiếp cận mới về kỹ thuật và lựa chọn tính năng cũng như nền tảng dữ liệu lớn. Khu vực dưới giá trị Đường cong ROC (AUC) đã được sử dụng để đánh giá hiệu suất của mô hình ML.

Tuy nhiên, các bài viết được đề cập ở trên đã không xem xét vấn đề mất cân bằng tập dữ liệu khi dự đoán sự rời bỏ của khách hàng ngân hàng. Trong bài báo này, chúng tôi muốn tập trung giải quyết vấn đề mất cân bằng tập dữ liệu bằng thuật toán RF dựa trên

SMOTE khi tạo dự đoán rời bỏ khách hàng trong lĩnh vực ngân hàng.

3. Phương pháp nghiên cứu

3.1. Thuật toán rừng ngẫu nhiên

RF là một trong những thuật toán học có giám sát phổ biến. RF được sử dụng cho các vấn đề của cả phân loại và hồi quy. RF là một cách tiếp cận tập hợp, bao gồm các cây quyết định khác nhau. RF cho thấy kết quả tốt hơn khi số lượng cây trong rừng tăng lên đáng kể và ngăn chặn việc khớp quá mức của mô hình. Mỗi kết quả được tạo ra từ mỗi cây quyết định trong rừng và những kết quả này được tích hợp để có được dự đoán chính xác và ổn định hơn [6].

Bộ dữ liệu được sử dụng trong bài viết này dựa trên Kaggle. Bộ dữ liệu bao gồm 13 đặc điểm (số hàng, id khách hàng, họ, kho tín dụng, địa lý, giới tính, tuổi, nhiệm kỳ, số dư, số lượng sản phẩm, hascrcard, thành viên đang hoạt động và lương ước tính) và một nhãn cho biết khách hàng có rời bỏ hay không.

***Bảng 1**. Mô tả dữ liệu*

Mô tả	*Chi tiết*
Số lượng cột/ Số lượng thuộc tính	13
Phân lớp (tỷ lệ rời bỏ)	Lớp 0: 7963 Lớp 1: 2037
Số lượng hàng	10000
Kiểu dữ liệu	float64(2), int64(9), object (3)
Dữ liệu bị thiếu	Không có

Dựa trên Bảng 1, chúng ta có thể nhận ra rằng tập dữ liệu được đề cập là tập dữ liệu không cân bằng. Để cải thiện độ chính xác của mô hình RF, các kỹ thuật lấy mẫu lại, được gọi là cấp độ thuật toán, cấp độ dữ liệu, học tập nhạy cảm với chi phí và dựa trên tập hợp được áp dụng [7]. Trong bài báo này, kỹ thuật lấy mẫu lại phổ biến, có tên là SMOTE được sử dụng để xử lý bộ dữ liệu mất cân bằng. Các kỹ thuật này phổ biến và đã chứng minh tính hiệu quả của chúng khi giải quyết các bộ dữ liệu mất cân bằng trong nhiều ứng dụng khác nhau [8], [9].

3.2. Các chỉ số đánh giá mô hình

Trong bài báo này, để đánh giá đầy đủ hiệu suất của mô hình ML, các số liệu đánh giá khác nhau, bao gồm độ chính xác, độ thu hồi, độ đặc hiệu, điểm F1, hệ số tương quan Mathews (MCC) và ROC-AUC được xem xét.

4. Kết quả

Kết quả của các chỉ số đánh giá RF dựa trên SMOTE bao gồm độ chính xác, thu hồi, điểm F1, ROC_AUC và điểm MCC được trình bày trong Bảng 2. Giá trị độ chính xác và ROC_AUC của RF dựa trên SMOTE là khá cao trong khi độ thu hồi, độ đặc hiệu, điểm F1, hệ số tương quan Mathews khá thấp. Điều này là do tập dữ liệu không được thu thập đầy đủ và tập dữ liệu tiền xử lý không được tối ưu hóa khi triển khai. Tuy nhiên, bài báo này sử dụng các thước đo đa dạng để đánh giá toàn diện mô hình học máy, đặc biệt là RF dựa trên SMOTE.

***Bảng 2**. Kết quả của các phương pháp đánh giá dự vào mô hình RF dựa trên SMOTE*

Chỉ số đánh giá	*Kết quả mô hình*
Độ chính xác (accuracy)	0.82
Độ thu hồi (recall)	0.66
Điểm F1 (F1-score)	0.6
ROC_AUC	0.75
Hệ số tương quan MCC	0.49

5. Kết luận

Trong bài viết này, chúng tôi sử dụng thuật toán RF dựa trên kỹ thuật lấy mẫu lại phổ biến SMOTE để tạo mô hình dự đoán về sự rời bỏ của khách hàng ngân hàng dựa trên bộ dữ liệu mất cân bằng thu được từ Kaggle. Các chỉ số được đánh giá khác nhau, bao gồm Độ chính xác, Thu hồi, Độ đặc hiệu, điểm F1, MCC và ROC-AUC được sử dụng để đo lường mô hình dự đoán rời bỏ khách hàng của ngân hàng nhằm đánh giá toàn diện. Thông qua các thử nghiệm của chúng tôi, tất cả các kết quả của số liệu đánh giá được chỉ ra trong bài viết này. Tuy nhiên, trong số các chỉ số đánh giá này, chỉ Độ chính xác và ROC-AUC hiển thị kết quả tích cực dựa trên tập dữ liệu được đề cập. Ngoài ra, để giúp

các ngân hàng phân tích tính năng nào ảnh hưởng nhiều nhất đến sự rời bỏ của khách hàng, bài viết này cũng xếp hạng các tính năng của tập dữ liệu từ điểm cao nhất đến điểm thấp nhất. Dựa trên phân tích này, các ngân hàng có chiến lược phù hợp để giữ chân khách hàng của mình.

Trong tương lai, hiệu suất của mô hình dự đoán rời bỏ khách hàng ngân hàng cần được cải thiện. Kết quả của các thước đo đánh giá này cũng đáng tin cậy hơn và đạt được giá trị cao. Để thực hiện được những mục tiêu đã nêu, trước tiên chúng ta cần thu thập chính xác bộ dữ liệu về sự rời bỏ của khách hàng trong lĩnh vực ngân hàng. Sau đó, việc sắp xếp lại dữ liệu và tiền xử lý phải được triển khai cẩn thận trên tập dữ liệu này để đạt được tập dữ liệu sạch và hiểu đầy đủ về tập dữ liệu này. Cuối cùng nhưng không kém phần quan trọng, một số thuật toán ML khác nhau dựa trên các kỹ thuật lấy mẫu lại khác nhau được sử dụng cho tập dữ liệu này để chọn mô hình dự đoán phù hợp nhất. Hơn nữa, một ứng dụng thực sự của việc rời bỏ khách hàng ngân hàng dựa trên các mô hình lý thuyết có thể được thực hiện trong tương lai.

Tài liệu tham khảo

[1] R. A. de Lima Lemos, T. C. Silva, and B. M. Tabak, “Propension to customer churn in a financial institution: a machine learning approach,” *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11751–11768, 2022, doi: 10.1007/s00521-022-07067-x.

[2] E. Villamosm, I. Kar, and D. Tansz, “Machine Learning Based Customer Decision Support,” pp. 1196–1201, 2020.

[3] “(Lecture Notes in Computer Science 12251) Osvaldo Gervasi, Beniamino Murgante, Sanjay Misra, Chiara Garau, Ivan Blečić, David Taniar, Bernady O. Apduhan, Ana Maria A. C. Rocha, Eufemia Tarantino, Carm.pdf.” .

[4] 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, 2022, doi: 10.1007/s00607-021-00908-y.

[5] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0191-6.

[6] T. K. Dang, T. C. Tran, and L. M. Tuan, “Machine Learning Based on Resampling Approaches and Deep Reinforcement Learning for Credit Card Fraud Detection Systems,” 2021.

[7] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Data Level Preprocessing Methods*. 2018.

[8] K. Li, W. Zhang, Q. Lu, and X. Fang, “An improved SMOTE imbalanced data classification method based on support degree,” *Proc. - 2014 Int. Conf. Identification, Inf. Knowl. Internet Things, IIKI 2014*, pp. 34–38, 2014, doi: 10.1109/IIKI.2014.14.

[9] T. C. Tran and T. K. Dang, “Machine Learning for Prediction of Imbalanced Data: Credit Fraud Detection,” *2021 15th Int. Conf. Ubiquitous Inf. Manag. Commun.*, pp. 1–7, 2021.