ỨNG DỤNG PHÂN LỚP DỮ LIỆU TRONG DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ TẠI NGÂN HÀNG THƯƠNG MẠI

Vũ Văn Hiệu[1*], Trương Hải Nam[2†]
[1]*Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và công nghệ Việt Nam,*
[2]*Học viện Chính sách và Phát triển*
Email:{vvhieu@ioit.ac.vn}, †{truonghainam66@gmail.com}*

Tóm tắt: *Bài toán giữ chân khách hàng không chỉ quan trọng với các ngân hàng còn là vấn đề ưu tiên của nhiều dịch vụ khác. Trong bài báo này, chúng tôi đề xuất mô hình học kết hợp (Stacking) cho dự đoán khách hành rời bỏ dịch vụ của ngân hàng sử dụng mô hình Stacking gồm hai cấp độ với cấp độ 0 gồm bốn mô hình cơ sở gồm K láng giềng gần nhất (KNN), XGBoost, rừng ngẫu nhiên (RDF) và máy hỗ trợ vector (SVM) và cấp độ một lần lượt gồm ba mô hình là hồi quy logistic, mạng neural hồi quy (RNN) và mạng neural học sâu (DNN). Phương pháp của chúng tôi đã thu được kết quả tốt với các chỉ số accuracy là 95,36%, recall là 95,57%, precision là 95,45% và F1 score là 95,51%.*

Từ khóa: Khách hàng rời bỏ, ngân hàng, học máy, học kết hợp.

Title: CUSTOMER DATA CLASSIFICATION APPLICATION FOR FORECASTING CUSTOMERS CHURN AT COMMERCIAL BANK

Abstract: The problem of customer churn is not only important for banks, but also a priority for many other services. The article proposes a Stacking model for predicting customer churn of banking services using a two-level Stacking model with Level 0 consisting of four basic models including K nearest neighbors (KNN), XGBoost, random forest (RDF) and support vector machine (SVM) and Level 1, respectively, include three models: logistic regression, regression neural network (RNN) and deep learning neural network (DNN). Our method has obtained good results with accuracy of 95.36%, recall of 95.57%, precision of 95.45% and F1 score of 95.51%.

Keywords: *Customer churm, banking, machine learning,* ensemble learning

---

1. Giới thiệu

Hiện nay, song hành cùng với sự phát triển ngày càng nhanh chóng của nền tảng công nghệ thông tin, có thể đại diện như là Blockchain, IOT, AI… thì các ngân hành thương mại cũng đồng thời nắm bắt với xu hướng phát triển chung cùng mọi lĩnh vực. Đi cùng với làn sóng chuyển đổi số trong ngân hàng thương mại hiện tại đã làm các ngân hành có sự thay đổi rất lớn và phức tạp, đồng thời cũng phải đối mặt với nhiều thách thức của thời cuộc. Cuộc chạy đua để giữ chân khách hàng giữa các ngân hàng thương mại cũng ngày càng thêm căng thẳng, mỗi ngân hàng đều đưa ra những chính sách riêng của họ nhưng để có thể cụ thể hóa những chính sách đó đến những khách hàng đang có ý định ngưng sử dụng dịch vụ là câu chuyện rất khác. Nếu các ngân hàng chỉ đợi chờ đến khi khách hàng đã rời bỏ dịch vụ của họ thì rất khó để có thể giữ chân họ tiếp tục với dịch vụ của mình. Điều này tạo cho ngân hàng một thách thức mới đó là phải dự đoán trước được những khách hàng nào sẽ rời bỏ dịch vụ trong tương lai. Với sự phát triển của ngành khoa học dữ liệu vào những năm gần đây giúp giải quyết được nhiều bài toán khác nhau đây là lựa chọn rất đúng đắn đối với các ngân hàng để giải quyết bài toán rời bỏ

dịch vụ của họ, với nguồn dữ liệu dồi dào từ khách hàng giúp việc dự đoán trở lên chính sác hơn. Bài nghiên cứu này ứng dụng machine learning vào việc dự đoán khách hàng rời bỏ làm căn cứ giúp các ngân hàng có một phương pháp để đưa ra các chính sách nhằm giữ chân khách hàng đến đúng với những khách hàng đang có ý định rời bỏ dịch vụ. Mục tiêu của bài nghiên cứu này nhằm mục đích tìm ra được phương pháp tốt nhất để có thể dự đoán được khả năng rời bỏ dịch vụ của khách hàng từ đó có thể giúp các ngân hàng có biện pháp kịp thời để giữ chân được khách hàng gắn bó với dịch vụ của mình.

Trong phần tiếp theo chúng tôi giới thiệu một số nghiên cứu liên quan ở Phần 2, Phương pháp đề xuất được đưa ra ở Phần 3. Cuối cùng là phần Thực nghiệm và kết quả, Kết luận trong Phần 4 và 5 tương ứng.

2. Nghiên cứu liên quan

Cho đến nay nhiều nhà nghiên cứu đã đề xuất các phương pháp nhằm giải quyết bài toán khách hàng rời bỏ bằng nhiều mô hình khác nhau. Trong [1] dự báo khách hàng rời bỏ dịch vụ của các ngân hàng thương mại dựa trên mô hình SVM. Nghiên cứu này sử dụng mô hình SVM, mô hình hồi quy logistic và mô hình RBF SVM. Kết quả đạt được với mô hình SVM và hồi quy logistic cho chỉ số accuracy đạt 99.09%, chỉ số recall bằng 0. Với mô hình RBF SVM cho ra kết quả dự đoán chính xác là 98,95% nhưng chỉ số recall là 26.84%, chỉ số precision là 39%. Từ các kết quả trên có thể thấy được nhóm tác giả đã không giải quyết được vấn đề mất cân bằng dữ liệu, điều này khiến mô hình không để dự đoán hoàn toàn các khách hàng rời bỏ dịch vụ.

Trong [2] đưa ra dự đoán khách hàng rời bỏ, nhóm tác giả đã sử dụng hai mô hình là KNN và Decision Tree (DT) để dự đoán khách hàng rời bỏ từ đó đưa ra so sánh để tìm ra mô hình có hiệu suất cao hơn. Nhóm tác giả đã đạt được kết quả với accuracy, precision, recall và F1 scose như sau: mô hình KNN cho kết quả lần lượt là 0.868, 0.615, 0.22 và 0.326, với mô hình DT cho ra kết quả lần lượt là 0.926, 0.775, 0.681 và 0.725. Qua các kết quả trên có thể thấy được cả hai thuật toán đều cho kết quả về độ chính xác đạt cao đối với KNN đạt 87%, và DT đạt 93%. Nhưng sự khác biệt lớn giữa hai thuật toán là ở F1 scose khi DT đạt 0.73, KNN chỉ đạt 0.33, nguyên nhân ở đây do là do sự chênh lệnh lớn ở recall khi mô hình KNN đạt rất thấp, điều này được nhóm tác giả giải thích do số lượng K láng giềng gây ra, khi cố cải thiện chỉ số recall bằng cách giảm số lượng K láng giềng.

Hệ thống dự đoán rời bỏ dịch vụ viễn thông trong [3] dựa trên việc học kết hợp bằng cách sử dụng nhóm đặc trưng. Nhóm tác giả đã xây dụng hệ thống mô hình Stacking và Voting với bốn thuật toán. Mô hình Stacking bao gồm hai cấp độ với Xgboost (XGB) ở cấp 1, và logistic regression (LR), decision tree (DT) và naive bayes classifier (NBC) ở cấp 2. Kết quả của hệ thống đề xuất cho thấy độ chính xác tốt nhất lần lượt là 96,12% và 98,09% đối với các bộ dữ liệu cũ và mới so với các hệ thống dự đoán khác.

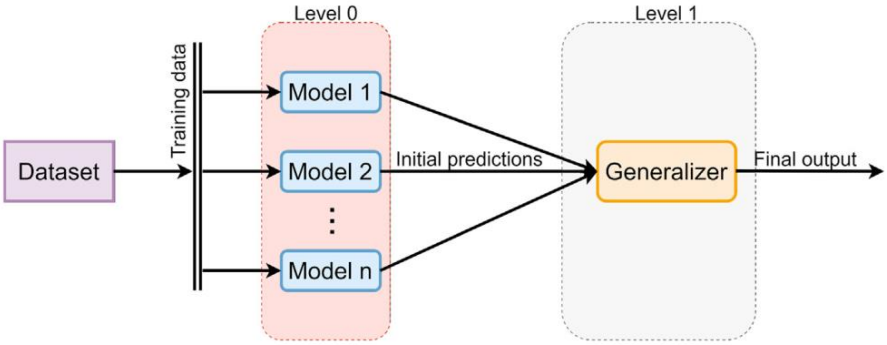
Dự đoán khách hàng rời bỏ của ngân hàng với học máy [4]. Nhóm tác giả đã xây dựng bao gồm 6 mô hình phân loại và tiến hành kiểm tra mô hình với kỹ thuật train/test split validation và Kfold cross validation và lần lượt các mô hình phân loại. Nghiên cứu đã đạt được độ chính xác lần lượt như sau: Logistic regression: 77.30%, Decision tree: 78,93%, Random forest: 87,22%, K-neighbors neighbors: 83,52%, AdaBoost: 83,55%, Gradient Boosting: 84,99% và XGBoost: 86.72%.

3. Phương pháp đề xuất

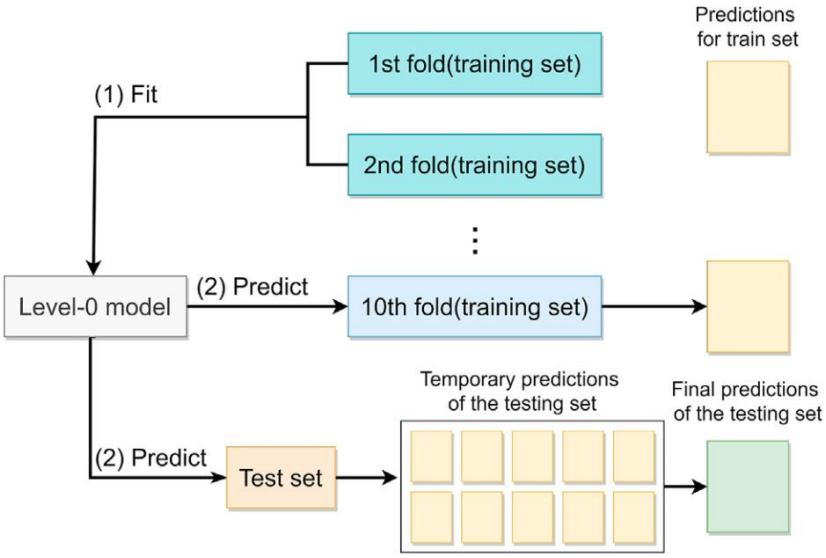
Nghiên cứu này đề xuất mô hình Stacking để dự đoán khách hàng rời bỏ của ngân hàng. Stacking là một phương pháp kết hợp nhiều mô hình cơ sở thành một siêu mô hình nhằm đạt được độ chính xác dự báo cao hơn [5] được giới thiệu trong [6]. Ở các mô hình cơ sở sẽ được huấn luyện trên tập dữ liệu huấn luyện, cuối cùng siêu mô hình sẽ được huấn luyện dựa trên các dự báo từ các mô hình cơ sở làm các đặc trưng. Ngoài nhiệm vụ giải quyết các bài toán phân loại phương pháp này cũng có thể sử dụng để hồi quy [7] và học không giám sát [8].

Hình 1 là mô hình Stacking gồm hai cấp độ với cấp độ 0 (level 0) gồm n mô hình được gọi là các mô hình cơ sở (Model Stack), ở cấp độ này các mô hình cơ sở có nhiệm vụ đưa ra các dự đoán làm thành đặc trưng cho mô hình ở cấp độ 1 (level). Ở cấp độ 1 sẽ gồm một mô

hình được gọi là siêu mô hình (Meta model), mô hình này có nhiệm vụ đưa ra các dự đoán cuối cùng từ các đặc trưng là các dự đoán của các mô hình cơ sở.



Hình 1: Phương pháp Stacking [9]



Hình 2: Biểu diễn sơ đồ mô hình cấp cơ sở [9]

Hình 2 gồm các mô hình cơ sở ở cấp độ 0 sẽ sử dụng phương pháp kiểm tra chéo (Cross-Validation) để có được các dự đoán từ làm các đặc trưng cho siêu mô hình. Phương pháp này nhằm mục đích lọc các kết quả đầu ra từ mức độ tổng quát đầu tiên. Việc huấn luyện hệ thống dự báo tiến hành như sau: N mẫu trong tập huấn luyện (S_i, y_i) với $1 \leq i \leq N$. Tập huấn luyện được chia ngẫu nhiên thành r đoạn với kích thước gần như bằng nhau để tạo thành $(S_i, y_i)_k$ trong đó $k' = (1, \dots r)$ là số đoạn. Lúc này $(S_i, y_i)_k$ thỏa mãn điều kiện như sau [9]:

$$\begin{cases} (S_i, y_i)_k \cup \overline{(S_i, y_i)_k} = (S, y) \\ (S_i, y_i)_k \cap \overline{(S_i, y_i)_k} = \emptyset \end{cases} \quad (1)$$

Đề xuất của chúng tôi áp dụng cho dự báo khách hàng rời bỏ dịch vụ dựa vào mô hình như trong Hình 2. Mô hình có cấp một (Level 0: Model Stack) là các mô hình cơ sở gồm bốn mô hình phân loại là K láng giềng gần nhất, XGBoost, rừng ngẫu nhiên và máy hỗ trợ vector. Ở cấp hai (Level 1: Meta Model) là các mô hình Hồi quy Logistic và Deep Neural Network (DNN) và Recurrent Neural Network (RNN). Cấu trúc mô hình đề xuất như Hình 3 và được thực hiện theo tiến trình theo Hình 4.

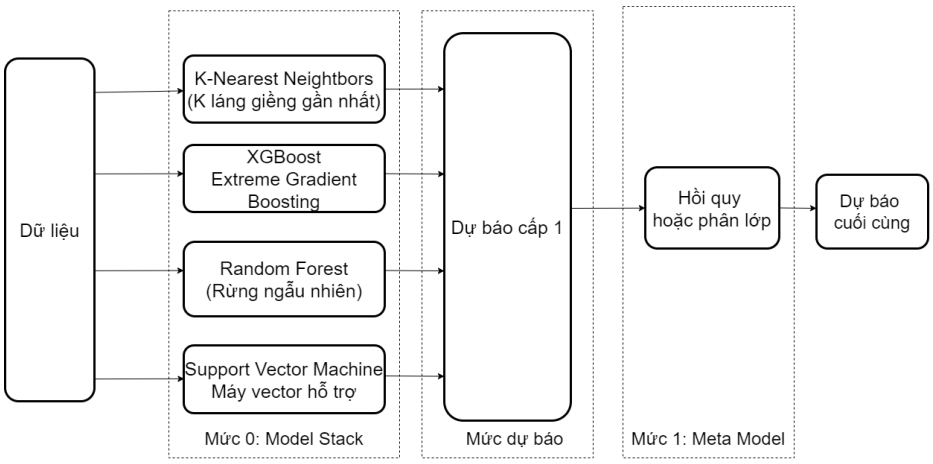
Tiến trình thực hiện bao gồm:

Bước 1: Khai thác dữ liệu

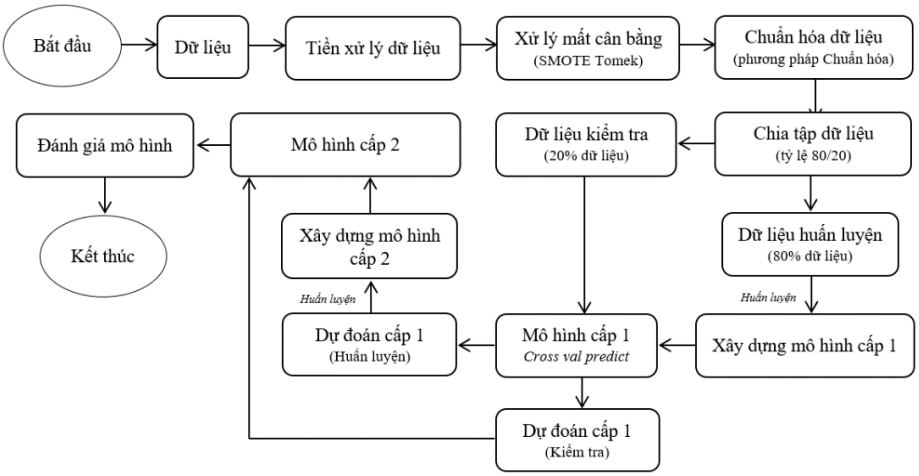
- Tiến hành xóa bỏ các trường dữ liệu không quan trọng, quan sát tổng quan dữ liệu (Hệ tương quan, giá trị thống kê, vấn đề mất cân bằng dữ liệu ở biến mục tiêu).

Bước 2: Chuẩn bị dữ liệu

- Tiến hành kiểm tra và xử lý dữ liệu (xóa trùng lặp, kiểm tra dữ liệu trống), biến đổi các trường dữ liệu định danh với phương pháp LabelEncoder.



Hình 3: Mô hình học Stacking



Hình 4: Biểu diễn tiến trình mô hình đề xuất

- Tiến hành chuẩn hóa dữ liệu với phương pháp Chuẩn hóa.
- Xử lý mất cân bằng dữ liệu với phương pháp SMOTE Tomek.

Bước 3: Xây dựng tập dữ liệu huấn luyện và kiểm tra

- Chia bộ dữ liệu thành 2 tập gồm tập huấn luyện 80% và tập dữ liệu kiểm tra 20%.

Bước 4: Xây dựng mô hình cấp 1

- Với tập dữ liệu huấn luyện tiến hành xây dựng các mô hình cấp 1 với phương pháp Grid search CV để tìm ra các siêu tham số.

Bước 5: Đưa ra dự đoán cấp 1

- Tập dữ liệu huấn luyện được chia thành 5 phần.
- Các mô hình cấp 1 đưa ra các dự đoán khi các mô hình sẽ học 4 phần là dự đoán phần còn lại, lập lại với toàn bộ các phần.
- Với các dự đoán của các mô hình cấp 1 được trở thành đặc trưng mới của tập dữ liệu huấn luyện.

Bước 6: Tập dữ liệu cấp 1 được huấn luyện để xây dựng Mô hình cấp 2.

Bước 7: Mô hình cấp 2 thử nghiệm với tập dữ liệu kiểm tra đưa ra đánh giá mô hình.

- Tập dữ liệu kiểm tra được các mô hình cấp 1 dự đoán làm thành các đặc trưng mới của tập dữ liệu kiểm tra.
- Đánh giá mô hình cấp 2 dự trên dự đoán của mô hình cấp 2 với tập dữ liệu kiểm tra.

3.1 Phương pháp đánh giá

Phương pháp đánh giá trong nghiên cứu này sử dụng các chỉ đánh giá accuracy, precision, recall và F1 scose, dựa trên ma trận nhầm lẫn có dạng như sau:

Accuracy là độ chính xác của mô hình, nó là tỷ lệ giữa số lượng mẫu được phân loại chính xác trên cho tổng số mẫu của tập dữ liệu thử nghiệm.

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

Precision là tỷ lệ giữa số lượng mẫu được dự đoán khớp mẫu trên tổng dự đoán khớp mẫu và dự đoán đúng nhưng sai. Precision được xác định như sau:

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

Recall là tỷ lệ giữa số lượng mẫu được dự đoán khớp mẫu với tổng dự đoán khớp mẫu và dự đoán sai nhưng đúng được xác định như sau:

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

Trong đó:

- TP (*True Positive*) là tổng số trường hợp dự báo khớp mẫu đúng,
- TN (*True Negative*) là tổng số trường hợp dự báo khớp với mẫu sai.
- FP (*False Positive*) là tổng số trường hợp dự báo các quan sát thuộc mẫu đúng tính thành sai.
- FN (*False Negative*) là tổng số trường hợp dự báo các quan sát thuộc mẫu sai tính thành đúng

F1 scose được hiểu là trung bình trọng số của Precision và Recall, nó thể hiện sự cân bằng giữa chúng, trong đó F1 scose đạt giá trị tốt nhất là 1 và kém nhất là 0, chỉ số F1 - scose được xác định như sau:

$$\text{F1 scose} = \frac{2 * \text{Precision} * Recall}{\text{Precision} + Recall}$$

4. Thực nghiệm và kết quả

4.1. Dữ liệu

Theo quy định pháp luật về việc bảo vệ thông tin cá nhân, vậy nên các ngân hàng thương mại sẽ không được quyền tiết lộ thông tin về tài khoản, gian dịch của khách hàng. Vì vậy trong bài nghiên cứu này sẽ thực nghiệm trên bộ dữ liệu *Churn_Modelling.csv* được tải xuống từ trang Kaggle.com [16] vào ngày 27/7/2022. Tập dữ liệu được tạo ra hoàn toàn không thuộc bất cứu ngân hàng nào, chỉ mang yếu tố mô tả dữ liệu có thể có trong ngân hàng. Bộ dữ liệu có mục đích nghiên cứu, thực nghiệm, trong bộ dữ liệu có 14 trường dữ liệu, gồm 10.000 quan sát và không có dữ liệu bị thiếu như trong Bảng 1. Trường dữ liệu *"Exited"* thể hiện khách hàng rời bỏ, trong đó có 2037 quan sát là rời bỏ được gán nhãn là '1' chiếm 20,37% và có 7963 quan sát là không rời bỏ được gán nhãn là '0' chiếm 79,63%. Bảng 2 là dữ liệu của 5 quan sát đầu tiên trong tập dữ liệu.

Bảng 1: Mô tả về bộ dữ liệu thực nghiệm

Trường dữ liệu	Mô tả	Kiểu dữ liệu
RowNumber	Số lượng khách hàng trong tập dữ liệu. tổng có 10.000 khách	int64
CustomerId	Mã khách hàng	int64
Surname	Tên khách hàng	Object
CreditScore	Điểm tín dụng đo lường mức tín nhiệm của khách hàng, điểm tín càng cao khánh hàng càng uy tín.	int64
Geography	Vị trí khách hàng sinh sống	Object
Gender	Giới tính (male – nam, female - nữ)	Object
Age	Tuổi	int64
Tenure	Khách hàng đã gắn bó với ngân hàng bao lâu	int64
Balance	Số dư trong tài khoản khách hàng	float64
NumOfProducts	Các sản phẩm mà khách hàng hiện đang sử dụng từ ngân hàng	int64
HasCrCard	Thể hiện khách hàng có thẻ tín dụng hay không (1 – có, 0 – không)	int64

IsActiveMember	Cho biết khách hàng có sử dụng bất kỳ sản phẩm nào của ngân hàng trong 6 tháng qua hay không (1 – có, 0 – không)	int64
EstimatedSalary	Mức lương ước tính của khách hàng	float64
Exited	Khách hàng có rời bỏ hay không (1 – có, 0 – không)	int64

4.2. Chuẩn bị dữ liệu

Chuẩn bị dữ liệu là quá trình lựa chọn, chuyển đổi dữ liệu đầu vào trước khi đưa vào quá trình huấn luyện mô hình, mục đích của quá trình này là giúp dữ liệu được đơn giản, tăng tốc độ chuyển đổi dữ liệu đồng thời nâng cao độ chính xác của mô hình. Kỹ thuật này là bắt buộc khi làm việc với các mô hình học máy [17]. Vấn đều đầu tiên cần được xử lý loại bỏ các trường dữ liệu không quan trọng trong quá trình huấn luyện mô hình. Trong (13) trường dữ liệu thuộc biến độc, trong đó có ba trường dữ liệu không quan trọng khi xây dựng mô hình, ba trường dữ liệu này bị loại bỏ đi bao gồm là RowNumber, CustomerId và Surname. Tiếp theo sẽ tiến hành biến đổi hai trường định danh là Geography và Gender bằng phương pháp LabelEncoder để chuyển thành dạng số. Các trường dữ liệu thuộc biến độc lập lúc này sẽ được chuẩn hóa Gaussian đưa dữ liệu về một phân bố trong phạm vi [-1, 1], chuẩn hóa được xác định như sau:

$$x' = \frac{x - \bar{x}}{\sigma} \qquad (2)$$

trong đó: x' là giá trị sau chuẩn hóa và $\bar{x}$ và σ lần lượt là giá trị trung bình và phương sai của các biến độc lập.

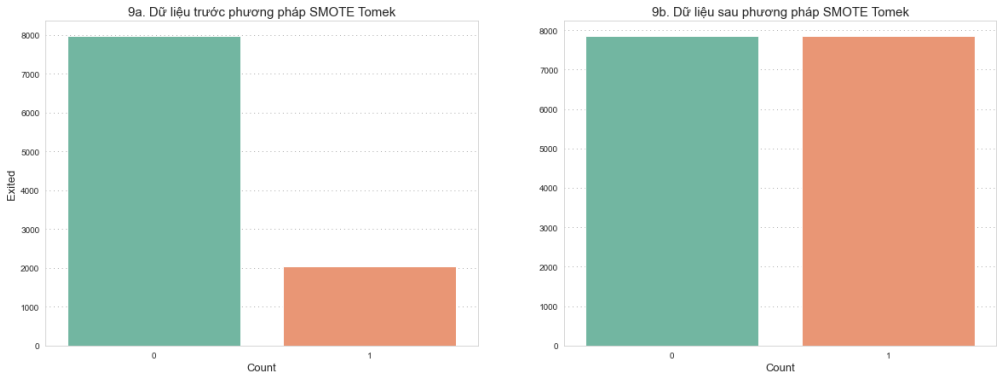
Bảng 4: Dữ liệu sau chuẩn hóa

0	1	2	3	4	5	6	7	8	9	Exited
-0.3262	-0.9019	-1.0960	0.2935	-1.0418	-1.2258	-0.9116	0.6461	0.9702	0.0219	1
-0.4400	1.5151	-1.0960	0.1982	-1.3875	0.1174	-0.9116	-1.5478	0.9702	0.2165	0
-1.5368	-0.9019	-1.0960	0.2935	1.0329	1.3331	2.5271	0.6461	-1.0307	0.2407	1

Vấn đề tiếp theo cần được giải quyết là sự mất cân bằng dữ liệu, có thể hiểu đơn giản mất cân bằng dữ liệu là một trong những trở ngại lớn trong bài toán phân loại. Các mô hình phân loại thường sẽ tập trung vào lớp đa số, do đó các mô hình không học tốt với lớp thiểu số. để đạt được độ chính xác cao hơn các điểm dữ liệu lớp thiểu số, nghĩa là số lượng phần tử đại diện cho một lớp lớn hơn rất nhiều so với lớp còn lại (lớp tối thiểu). Trong tập dữ liệu thực nghiệm có thể thấy khách hàng rời bỏ chỉ chiếm 20,37% so với khách hàng không rời bỏ là 79,63%, tỷ lệ giữa khách hàng không rời bỏ dịch vụ cao hơn rất nhiều so với khách hàng rời bỏ (Hình 9a).

Có nhiều phương pháp giải quyết vấn đề này với hướng tiếp cận ở mức độ dữ liệu, bao gồm các phương pháp điều chỉnh để giảm sự mất cân bằng dữ liệu bằng cách giảm số lượng phần tử ở lớp đa số hoặc tăng số lượng phần tử tối thiểu (sinh thêm một cách ngẫu nhiên hoặc sinh thêm phần tử nhân tạo). Bởi vì hạn chế về thích thước của bộ dữ liệu thực nghiệm vậy nên để giải quyết vấn đề mất cân bằng dữ liệu, bài sẽ sử dụng phương pháp SMOTE Tomek là phương pháp kết hợp khả năng của SMOTE là sinh thêm phần tử thiểu số nhân tạo và khả năng của Tomek Links để xóa dữ liệu được xác định là liên kết Tomek khỏi lớp đa số, có nghĩa là các mẫu dữ liệu từ lớp đa số gần nhất với dữ liệu lớp thiểu số (Hình 9b).

Hình 9: Dữ liệu trước và sau phương pháp SMOTE Tomek



Toàn bộ dữ liệu sẽ được chia thành hai tập dữ liệu là tập huấn luyện (80%) và tập thử nghiệm (20%). Việc chia hai tập dữ liệu theo tỷ lệ 80:20 giúp mô hình đạt được hiệu quả tốt vì tỷ lệ dữ liệu huấn luyện hơn dẫn đến phương sai lớn hơn trong các ước tính tham số của mô hình, trong khi ít dữ liệu thử nghiệm hơn dẫn đến phương sai lớn hơn trong hiệu suất của mô hình. Vì vậy mục tiêu là để đảm bảo rằng dữ liệu được chia thành các tập huấn luyện và thử nghiệm dẫn đến một phương sai không quá cao, vậy nên sẽ lựa chọn tỷ lệ 80:20 [18].

4.3. Kết quả

Trong mô hình đề xuất để các mô hình cơ sở (Mức 0) có thể dự báo, mô hình cơ sở cần huấn luyện với tập dữ liệu huấn luyện, từ đó có thể có thể đánh giá được các kết quả dự báo từ các mô hình này. Bảng 5 là các chỉ số đánh giá hiệu năng trên một số bộ phân loại KNN, RDF, SVM và XGB.

Bảng 5: Kết quả đánh giá các mô hình cấp 1

	KNN	**RDF**	**SVM**	**XGB**
Accuracy	0.8751	0.9092	**0.9286**	0.9051
Recall	0.9582	0.9077	**0.9631**	0.8733
Precision	0.8278	0.9162	0.9052	**0.9394**
F1-score	0.8882	0.9119	**0.9332**	0.9050

Bảng 6: Kết quả đánh giá mô hình cấp 2

	ML-LR (Stacking – Logistic regression)	**ML-DNN (Stacking – DNN)**	**ML-RNN (Stacking – RNN)**
Accuracy	**0.9534**	0.9346	0.9334
Recall	**0.9557**	0.9151	0.9255
Precision	0.9545	**0.9569**	0.9447
F1-score	**0.9551**	0.9355	0.9350
Thời gian	**2 phút 56s**	7 phút 55s	6 phút 24s

Bảng 6 là kết quả so sánh các chỉ số hiệu năng của mô hình đề xuất ở Mức 2 khi sử dụng lần lượt các bộ phân loại khác nhau (LR, DNN, RNN). Kết quả cho thấy mô hình ML-LR có kết quả tốt nhất khi tất cả các chỉ số đều đạt trên 95% và các chỉ số accuracy, recall và F1-score đều cao hơn hai mô hình còn lại từ 2 đến 4%, ngoại trừ chỉ số precison thấp hơn mô hình ML-DNN chỉ là 0,024% và thời gian huấn luyện nhanh hơn rất nhiều so với hai mô hình còn lại là ML-DNN với 5 phút và ML-RNN là 3 phút 28 giây.

Bảng 7: So sánh kết quả của mô hình đề xuất với các nghiên cứu trên tập dữ liệu [16]

	ML-LR	DNN[19]	XGB*[20]	KNN*[20]
Accuracy	**0.9534**	0.8590	0.8685	0.8385
Recall	0.9557	0.9492	**0.9673**	0.9516
Precision	**0.9545**	0.8827	0.8795	0.8603
F1-score	**0.9551**	0.9147	0.9213	0.9036

Trong Bảng 7 so sánh mô hình đề xuất với các mô hình trong một số nghiên cứu khác trên cùng tập dữ liệu [16]. Mô hình đề xuất có hiệu suất đánh giá cao khi các chỉ số đánh giá đều ở mức 95%, cho thấy mô hình có sự ổn định và không có sự chênh lệnh giữa các chỉ số đánh giá. Mô hình đề xuất có chỉ số accuracy là 95% cao hơn các mô hình còn lại lần lượt là 10% DNN[19] , 9% XGB*[20] và 12% KNN*[20], hai chỉ số là precision và F1-score đều cao hơn từ 3% đến 9% so với các mô hình khác. Chỉ số recall trong cả bốn mô hình đều có chỉ số cao nằm trong mức từ 94 đến 96%.

5. Kết luận

Nghiên cứu này cung cấp cho các ngân hàng thương mại một công cụ giúp có thể dự đoán được khách hàng rời bỏ dịch vụ, nhằm giúp cải thiện được khả năng giữ chân khách hàng bằng đề xuất sử dụng mô hình Stacking. Mô hình đề xuất sử dụng kết hợp nhiều mô hình học máy giúp được hiệu suất của mô hình. Mô hình đề xuất gồm hai cấp độ, cấp độ một gồm bốn mô hình cơ sở là KNN, SVM, XGB và RDF, duy nhất một mô hình cấp độ hai có thể là LR, DNN hoặc RNN. Việc kết hợp nhiều mô hình thành một giúp tăng đáng kể khả năng phát hiện khách hàng rời bỏ, khi mô hình đề xuất thu được các chỉ số đánh giá cao nhất với mô hình Stacking với Mức hai là mô hình Logistic regression khi đạt được accuracy là 95.34% và F1-score là 95.51%.

TÀI LIỆU THAM KHẢO

[1]. He, Benlan et al. *"Prediction of Customer Attrition of Commercial Banks based on SVM Model."* ITQM (2014).

[2]. Hassonah, Mohammad A. et al. *"Churn Prediction: A Comparative Study Using KNN and Decision Trees."* 2019 Sixth HCT Information Technology Trends (ITT) (2019): 182-186.

[3]. Tianpei Xu, Ying Ma, Kangchul Kim (2021), "*Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping.*" Department of Computer Engineering, Chonnam National Unversity, Yeosu 59626, Korea.

[4]. Kumar, S. Likhit. *"Bank Customer Churn Prediction Using Machine Learning."* International Journal for Research in Applied Science and Engineering Technology (2021): n. pag.

[5]. Ting, Kai Ming and Ian H. Witten. *"Issues in Stacked Generalization*." J. Artif. Intell. Res. 10 (1999): 271-289.

[6]. Wolpert, David H.. *"Stacked generalization."* Neural Networks 5 (1992): 241-259.

[7]. Breiman, L.. *"Stacked regressions."* Machine Learning 24 (1996): 49-64.

[8]. Smyth, Padhraic and David H. Wolpert. *"Stacked Density Estimation."* NIPS (1997).

[9]. Massaoudi, Mohamed et al. *"A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting."* Energy (2021): n. pag.

[10]. Altman, Naomi S.. *"An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression."* The American Statistician 46 (1992): 175-185.

[11]. Tianqi Chen, Carlos Guestrin (2016), *"XGBoost: A Scalable Tree Boosting System."* In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[12]. Müller, Andreas and Sarah Guido. *"Introduction to Machine Learning with Python: A Guide for Data Scientists."* (2016).

[13]. Kanchan Sarkar . "*ReLU : Not a Differentiable Function: Why used in Gradient Based Optimization? and Other Generalizations of ReLU*". (2018) Nguồn: https://medium.com/@kanchansarkar/relu-not-a-differentiable-function-why-used-in-gradient-based-optimization-7fef3a4cecec (truy cập 3/8/2022).

[14]. Bayraci, Selçuk and Orkun Susuz. *"A Deep Neural Network (DNN) based classification model in application to loan default prediction."* Theoretical and Applied Economics (2019): 75-84.
[15]. R. Collobert. "*Large Scale Machine Learning*". PhD thesis, Universite de Paris VI, LIP6, 2004.
[16] Bank Turnover Dataset.https://www.kaggle.com/datasets/barelydedicated/bank-customer-churn-modeling.
[17]. Harshil Patel (2021), "*What is Feature Engineering — Importance, Tools and Techniques for Machine Learning*". Nguồn: https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning2080b0269f10 (truy cập 2/8/2022).
[18]. Sina E. Charandabi (2020). "*Prediction of Customer Churn in Banking Industry*." Department of Decision Science, LeBow College of Business, Drexel University.
[19]. Domingos, Edvaldo et al. *"Experimental Analysis of Hyperparameters for Deep Learning-Based Churn Prediction in the Banking Sector."* Comput. 9 (2021): 34.
[20]. Hemlata Dalmia, Ch V S S Nikil, Sandeep Kumar. *"Churning of Bank Customers Using Supervised Learning."* Innovations in Electronics and Communication Engineering (2020) (pp.681-691).