Ứng dụng PCA kết hợp mô hình SVM xây dựng mô hình xích Markov lai và mô phỏng Monte Carlo tối ưu chiến lược giữ chân khách hàng

Trần Minh Phát*, Nguyễn Đình Tuấn Phúc*, Nguyễn Bảo Quân*
*Khoa Khoa học và Kĩ thuật Thông Tin
Trường Đại Học Công nghệ Thông tin, Đại học Quốc gia TP. Hồ Chí Minh
Email: {24521317@gm.uit.edu.vn, 24521384@gm.uit.edu.vn, 24521436@gm.uit.edu.vn }

***Tóm tắt nội dung*—Dự đoán khách hàng rời bỏ dịch vụ là yếu tố quan trọng giúp doanh nghiệp tối ưu chiến lược giữ chân khách hàng và giảm thiệt hại doanh thu. Bài báo cáo đề xuất khung phương pháp kết hợp Phân tích Thành phần Chính (PCA) và Máy Vector Hỗ trợ (SVM) để đánh giá rủi ro rời bỏ cá nhân. Bộ dữ liệu ban đầu gồm 23 đặc trưng được giảm xuống còn 18 thành phần chính, giữ 96,21% phương sai tổng thể, đồng thời bảo toàn thông tin quan trọng. Hành vi chuyển trạng thái khách hàng được mô hình hóa bằng xích Markov lai, kết hợp mô phỏng Monte Carlo với 1.000 lần lặp trên 30 khách hàng mẫu, phản ánh sự phân tán rủi ro cá nhân. Kết quả thực nghiệm đạt Accuracy 0,847410, Brier Score 0,1264 và cải thiện nhẹ AUC, chứng minh phương pháp dự đoán xác suất rời bỏ hiệu quả, hỗ trợ doanh nghiệp ra quyết định dựa trên dữ liệu.**

***Index Terms*—Dự đoán rời bỏ khách hàng, PCA, Học Máy Vector Hỗ trợ, Xích Markov lai, Mô phỏng Monte Carlo.**

I. Giới Thiệu

TRONG kỷ nguyên số hóa, thị trường viễn thông đang đối mặt với mức độ cạnh tranh ngày càng khốc liệt khi số lượng người dùng mới dần bão hòa. Do đó, việc giữ chân khách hàng hiện tại trở thành yếu tố then chốt, bởi chi phí thu hút khách hàng mới thường cao hơn gấp 5 đến 25 lần so với chi phí duy trì khách hàng cũ [1]. Hiện tượng khách hàng rời bỏ dịch vụ (*Customer Churn*) không chỉ làm suy giảm doanh thu mà còn ảnh hưởng trực tiếp đến lợi thế cạnh tranh dài hạn của doanh nghiệp [2].

Sự phát triển của dữ liệu lớn tạo điều kiện cho việc phân tích hành vi và lịch sử giao dịch của khách hàng, nhưng đồng thời đặt ra thách thức về xử lý các tập dữ liệu có số chiều lớn và độ nhiễu cao [3]. Nhằm giải quyết vấn đề này, nghiên cứu đề xuất một khung phương pháp kết hợp Phân tích Thành phần Chính (*PCA*) và Máy Vector Hỗ trợ (*SVM*) để đánh giá tác động của giảm chiều dữ liệu đến hiệu suất tính toán và độ chính xác dự đoán [4]. Bên cạnh đó, hành trình chuyển đổi trạng thái của khách hàng được mô hình hóa bằng xích *Markov*, kết hợp mô phỏng *Monte Carlo* nhằm dự báo các kịch bản *churn* trong tương lai, giúp khắc phục hạn chế của các mô hình phân loại tĩnh truyền thống [5].

Các đóng góp chính của bài báo được tóm tắt như sau:

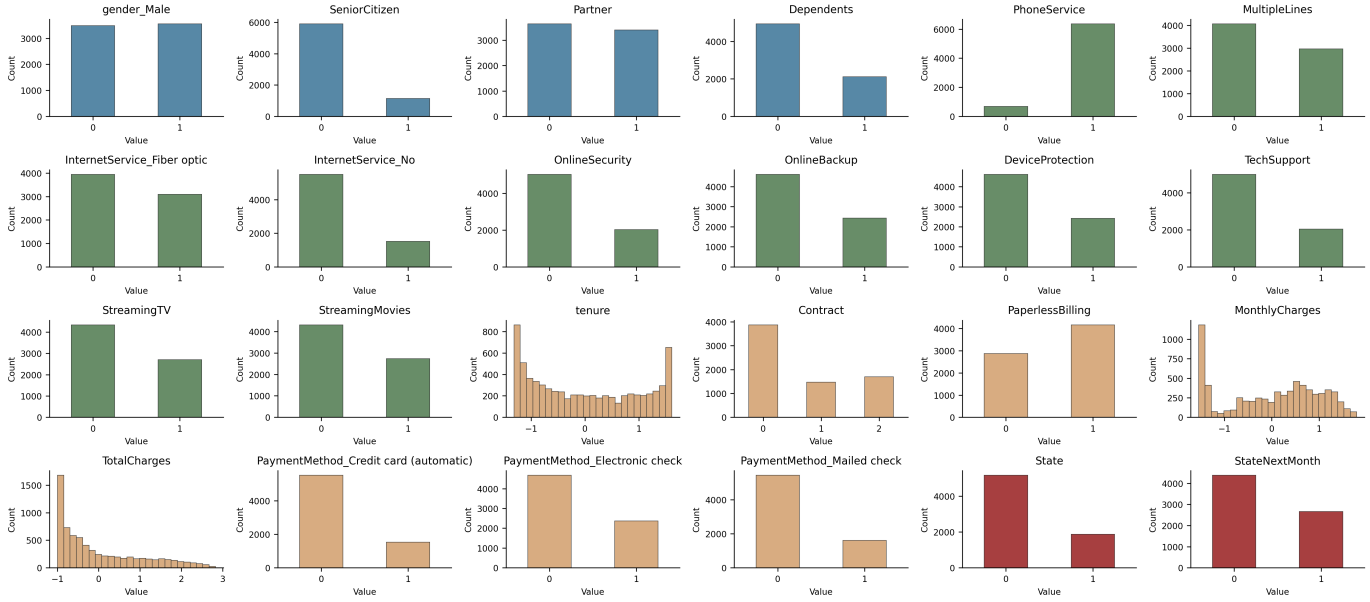
- **Tối ưu hóa Mô hình Dự báo:** Đề xuất mô hình kết hợp PCA và SVM nhằm tinh gọn không gian vector đặc trưng, loại bỏ nhiễu và giảm thiểu đáng kể chi phí tính toán trong khi vẫn duy trì độ chính xác cao.
- **Mô hình hóa Động lực Khách hàng:** Xây dựng mô hình xích Markov lai để mô hình hóa sự chuyển dịch trạng thái khách hàng theo thời gian, cho phép xác định xác suất chuyển đổi động giữa các nhóm "churn" và "no churn", thay dừng lại ở dự báo tĩnh tại một thời điểm.
- **Dự báo và Tối ưu Chiến lược:** Tích hợp mô phỏng Monte Carlo để giả lập các biến động thị trường dài hạn, cung cấp các kịch bản dự báo định lượng về quy mô khách hàng, hỗ trợ doanh nghiệp tối ưu hóa chiến lược giữ chân và phân bổ ngân sách tiếp thị hiệu quả.

II. Các Công Trình Liên Quan

Nghiên cứu dự báo *churn* đã chuyển dịch mạnh mẽ từ các mô hình thống kê truyền thống sang các thuật toán học máy hiện đại với độ phức tạp cao hơn [1], [2].

Học máy và giảm chiều dữ liệu: Mô hình **SVM** được đánh giá cao nhờ khả năng tối ưu hóa siêu phẳng trong không gian đa chiều [6], nhưng thường nhạy cảm với nhiễu dữ liệu. Phương pháp **PCA** thường được áp dụng song song để giảm gánh nặng tính toán và tăng độ chính xác thông qua việc loại bỏ các biến tương quan [4]. Tuy nhiên, việc kiểm soát ngưỡng giữ lại thành phần chính là yếu tố sống còn để tránh mất mát các thông tin đặc thù liên quan đến hành vi khách hàng rời bỏ [7].

Mô hình hóa trạng thái và mô phỏng: Bên cạnh các dự báo tĩnh, **xích Markov** giúp mô hình hóa sự chuyển dịch lòng trung thành và tính toán giá trị vòng đời khách hàng (*CLV*) [5]. Nhằm khắc phục tính xác định (*deterministic*) của mô hình *Markov* thuần túy, mô phỏng **Monte Carlo** được tích hợp để giả lập các kịch bản ngẫu nhiên. Cách tiếp cận này cho phép doanh nghiệp đánh giá rủi ro và biên độ hiệu quả của các chiến dịch giữ chân một cách định lượng dưới tác động của tính bất định trong hành vi người dùng [8].



Hình 1. Phân phối giá trị các thuộc tính của tập dữ liệu sau khi tiền xử lí.

III. Mô tả Dữ liệu và Tiền xử lý

Nghiên cứu sử dụng bộ dữ liệu TelcoChurn được công bố trên nền tảng Kaggle, bao gồm **7.043 khách hàng với 23 thuộc tính**. Nhãn mục tiêu của bài toán là *StateNextMonth*, biểu diễn trạng thái rời bỏ (*Churn*) hoặc tiếp tục sử dụng dịch vụ (*No Churn*) của khách hàng trong tháng kế tiếp. Các thuộc tính được phân thành các nhóm chính gồm: thông tin nhân khẩu học, thông tin sử dụng dịch vụ, thông tin hợp đồng và phương thức thanh toán, thông tin chi phí dịch vụ và trạng thái hiện tại của khách hàng.

Quy trình tiền xử lý dữ liệu được thực hiện theo các bước sau:

A. Làm sạch dữ liệu và chuyển đổi kiểu dữ liệu

Thuộc tính định danh *customerID* được loại bỏ do không mang thông tin dự đoán. Thuộc tính *TotalCharges* được chuyển đổi từ dạng chuỗi sang dạng số.

B. Mã hóa các thuộc tính phân loại nhị phân

Tất cả các thuộc tính dạng *Yes/No*, bao gồm cả các thuộc tính liên quan đến dịch vụ (*MultipleLines*, *OnlineSecurity*, *OnlineBackup*, *DeviceProtection*, *TechSupport*, *StreamingTV* và *StreamingMovies*), được thống nhất mã hóa về dạng nhị phân. Các giá trị đặc biệt như *No internet service* và *No phone service* được quy về nhóm *No* trước khi mã hóa.

C. Mã hóa thuộc tính có thứ tự

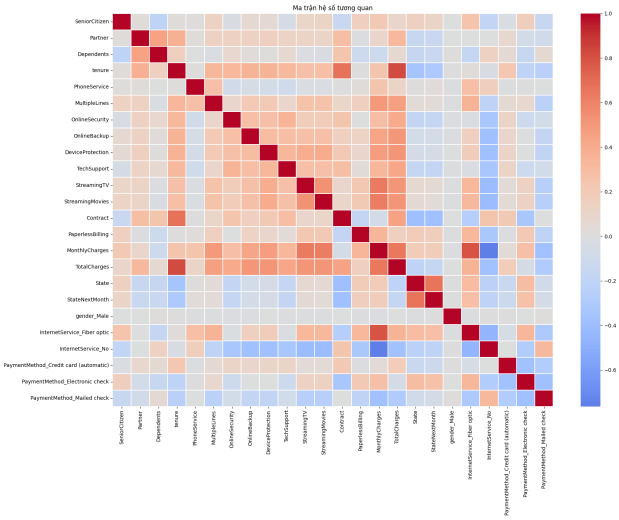
Thuộc tính *Contract* được mã hóa theo thứ tự tăng dần về mức độ cam kết của khách hàng (*ordinal encoding*), lần lượt tương ứng với các giá trị *Month-to-month*, *One year* và *Two year*.

D. Mã hóa nhãn trạng thái

Hai thuộc tính *State* và *StateNextMonth* được ánh xạ sang dạng nhị phân, trong đó *Churn* được gán giá trị 1 và *No Churn* được gán giá trị 0 (*label encoding*).

E. Mã hóa one-hot cho các thuộc tính phân loại không có thứ tự

Các thuộc tính *gender*, *InternetService* và *PaymentMethod* được xử lý bằng phương pháp *one-hot encoding*, đồng thời loại bỏ cột giả đầu tiên nhằm tránh hiện tượng đa cộng tuyến (*multicollinearity*).



Hình 2. Ma trận hệ số tương quan

F. Chuẩn hóa dữ liệu số

Các thuộc tính số gồm *tenure*, *MonthlyCharges* và *TotalCharges* được chuẩn hóa về phân phối chuẩn bằng phương pháp *StandardScaler* (*z-score normalization*) giúp đảm bảo các đặc trưng đóng góp cân bằng vào quá trình huấn luyện, đặc biệt đối với các mô hình dựa trên khoảng cách và biên như Support Vector Machine (*SVM*).

Sau quá trình tiền xử lý, bộ dữ liệu thu được hoàn toàn ở dạng số, không chứa giá trị thiếu và đã được chuẩn hóa về miền giá trị cũng như phân phối. Phân phối các thuộc tính và ma trận tương quan sau tiền xử lý được trình bày lần lượt trong Hình 1 và Hình 2.

IV. Phương Pháp Đề Xuất

Khung phương pháp đề xuất tích hợp PCA, SVM, xích Markov lai và mô phỏng Monte Carlo nhằm tối ưu hóa dự báo rủi ro rời bỏ như Hình 3.

A. Giảm chiều dữ liệu bằng Phân tích thành phần chính (PCA)

Principal Component Analysis (*PCA*) được áp dụng nhằm giải quyết hiện tượng đa cộng tuyến (*multicollinearity*) giữa các biến số, đồng thời giảm chiều dữ liệu. Cụ thể, *PCA* thực hiện phép biến đổi tuyến tính để chiếu dữ liệu ban đầu sang một không gian trực giao mới, trong đó các thành phần chính là độc lập tuyến tính với nhau và được sắp xếp theo thứ tự giảm dần của phương sai.

Nhóm sẽ lựa chọn số lượng thành phần chính k dựa trên ngưỡng phương sai tích lũy, nhằm đảm bảo duy trì phần lớn thông tin của dữ liệu gốc trong khi giảm thiểu số chiều. Cụ thể, giá trị k được xác định sao cho tổng phương sai được giải thích bởi k thành phần chính đầu tiên đạt ít nhất **95%** tổng phương sai của toàn bộ dữ liệu, theo công thức:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{D} \lambda_j} \geq 0.95 \tag{1}$$

trong đó λ_i là trị riêng (*eigenvalue*) tương ứng với thành phần chính thứ i, phản ánh lượng phương sai được giải thích bởi thành phần đó, và D là số chiều ban đầu của không gian dữ liệu.

B. Mô hình Support Vector Machine (SVM)

SVM là một mô hình học có giám sát dựa trên nguyên lý tối đa hóa biên (*maximum margin*), trong đó mục tiêu là tìm một siêu phẳng phân tách sao cho khoảng cách đến các điểm dữ liệu gần nhất của hai lớp là lớn nhất. Đối với bài toán phân loại nhị phân, hàm quyết định của **SVM** được biểu diễn dưới dạng:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \tag{2}$$

trong đó: α_i là các hệ số *Lagrange*, $y_i \in \{-1, +1\}$ là nhãn lớp, và b là hệ số chệch.

Nhằm nắm bắt các ranh giới quyết định phi tuyến (*nonlinear decision boundaries*) trong bài toán dự đoán hành vi rời bỏ khách hàng (*churn prediction*), nhóm sử dụng hàm nhân **Gaussian Radial Basis Function** (*RBF*), được định nghĩa như sau:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{3}$$

trong đó: x_i, x_j là các vector đặc trưng trong không gian, và $\gamma > 0$ là siêu tham số *kernel*.

Đầu ra gốc của mô hình **SVM** là giá trị hàm quyết định (*decision function*) $f(x)$, biểu diễn khoảng cách có dấu từ điểm dữ liệu đến siêu phẳng phân tách. Giá trị này nằm trong miền:

$$f(x) \in (-\infty, +\infty) \tag{4}$$

và không mang ý nghĩa xác suất trực tiếp (*no direct probabilistic meaning*), do đó không phù hợp cho các bài toán đánh giá và ra quyết định dựa trên xác suất, chẳng hạn như phân tích rủi ro hoặc mô hình *Markov* trong dự đoán *churn*.

Để chuyển đổi đầu ra của **SVM** sang dạng xác suất hậu nghiệm (*posterior probability*), nghiên cứu áp dụng **kỹ thuật Platt scaling** (*Platt scaling*). Phương pháp này khớp một hàm *sigmoid* (*sigmoid function*) lên các giá trị $f(x_i)$ của **SVM**, theo công thức:

$$P(y = \text{Churn} \mid x_i) = \frac{1}{1 + \exp\left(Af(x_i) + B\right)} \tag{5}$$

trong đó: A và B là các tham số được ước lượng thông qua tối ưu hóa *log-likelihood* (*log-likelihood optimization*).

C. Xích Markov Lai

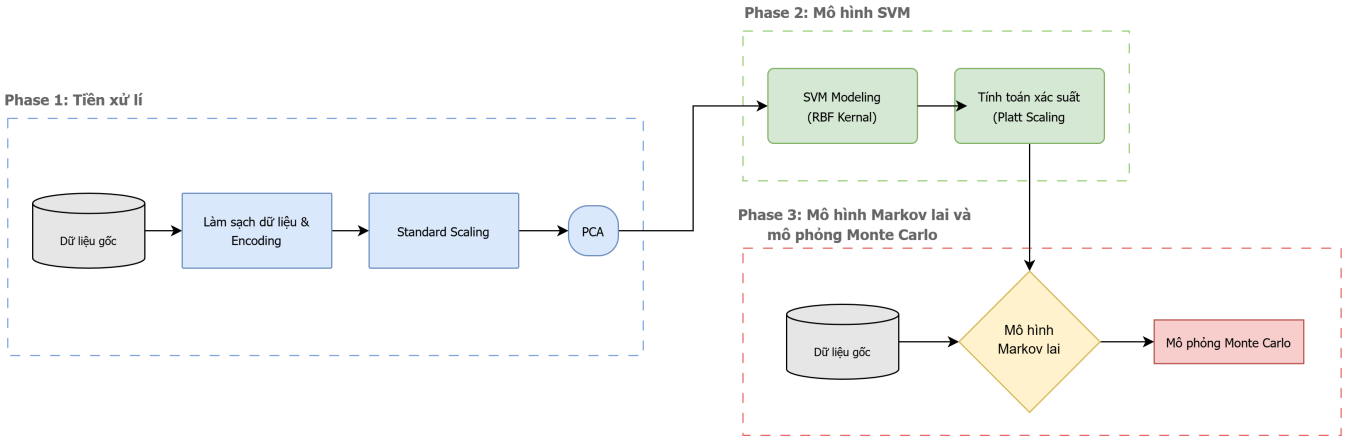
Xích Markov lai: Hành vi của khách hàng được mô hình hóa như một xích *Markov* rời rạc theo thời gian $\{S_t\}_{t \geq 0}$, thỏa mãn tính chất *Markov* (*Markov property*):

$$\mathbb{P}\left(S_t = j \mid S_{t-1} = i, S_{t-2}, \ldots, S_0\right) = \mathbb{P}\left(S_t = j \mid S_{t-1} = i\right) \tag{6}$$

Trong bài toán *churn*, mỗi khách hàng tại thời điểm t thuộc một trong hai trạng thái: $S_t \in \{0, 1\}$, trong đó $S_t = 0$ biểu thị khách hàng đang hoạt động (*No Churn*) và $S_t = 1$ biểu thị khách hàng đã rời bỏ (*Churn*). Do dữ liệu về hành vi sau khi khách hàng rời bỏ thường hạn chế, nhóm đề xuất sử dụng **xích Markov lai** (*hybrid Markov chain*), trong đó các xác suất chuyển trạng thái được xác định theo hai cơ chế khác nhau. Với mỗi khách hàng i, ma trận chuyển trạng thái (*transition matrix*) được định nghĩa như sau:

$$\mathbf{M}^{(i)} = \begin{bmatrix} 1 - p_i & p_i \\ \alpha & 1 - \alpha \end{bmatrix} \tag{7}$$

Trong đó: $p_i = \mathbb{P}\left(S_t = 1 \mid S_{t-1} = 0, i\right)$ là xác suất khách hàng i rời bỏ ở bước tiếp theo, được lấy từ đầu ra xác suất của mô hình **PCA–SVM** sau khi hiệu chỉnh bằng **Platt scaling** nhằm đảm bảo tính hiệu chuẩn xác suất (*probability calibration*); $\alpha = \mathbb{P}\left(S_t = 0 \mid S_{t-1} = 1\right)$ là xác suất khách hàng quay lại sau khi đã rời bỏ, được ước lượng từ thống kê thực nghiệm trên toàn bộ tập dữ liệu lịch sử. Cách xây dựng này cho phép mô hình phản ánh mức độ rủi ro riêng của



Hình 3. Sơ đồ phương pháp tích hợp PCA, SVM, xích Markov lai và mô phỏng Monte Carlo nhằm tối ưu hóa dự báo rủi ro rời bỏ.

từng khách hàng, đồng thời tránh hiện tượng ước lượng không ổn định tại trạng thái *Churn* do thiếu dữ liệu quan sát.

Vector xác suất trạng thái tại thời điểm t được xác định thông qua công thức:

$$\boldsymbol{\pi}_t = \boldsymbol{\pi}_0 \cdot (\mathbf{P}^{(i)})^t \tag{8}$$

trong đó $\boldsymbol{\pi}_0 = (1, 0)$ biểu diễn trạng thái ban đầu của khách hàng là *No Churn*, và $\mathbf{P}^{(i)}$ là ma trận chuyển trạng thái tương ứng với khách hàng thứ i.

D. Mô phỏng Monte Carlo

Mô phỏng **Monte Carlo** là một kỹ thuật xấp xỉ số (*numerical approximation*), trong đó các đại lượng kỳ vọng được ước lượng thông qua việc lấy mẫu ngẫu nhiên lặp lại từ một quá trình xác suất đã biết. Với một biến ngẫu nhiên X, kỳ vọng toán học của X được xấp xỉ bởi:

$$\mathbb{E}[X] \approx \frac{1}{N} \sum_{k=1}^{N} X^{(k)} \tag{9}$$

trong đó $X^{(k)}$ là giá trị của biến ngẫu nhiên trong lần mô phỏng thứ k và N là số lượng mô phỏng. Theo **định luật số lớn** (*Law of Large Numbers*), xấp xỉ này hội tụ về giá trị kỳ vọng thực khi N đủ lớn.

Trong bối cảnh xích *Markov*, mô phỏng **Monte Carlo** cho phép sinh ra nhiều quỹ đạo trạng thái (*state trajectories*) khác nhau của quá trình $\{S_t\}$, từ đó ước lượng phân phối và kỳ vọng của các đại lượng quan tâm theo thời gian. Thay vì chỉ sử dụng phép nhân ma trận xác suất để tính phân phối trạng thái, nghiên cứu áp dụng mô phỏng **Monte Carlo** nhằm: (i) nắm bắt tính ngẫu nhiên trong hành vi khách hàng; (ii) dễ dàng tích hợp các đại lượng phụ thuộc trạng thái, chẳng hạn như doanh thu hoặc giá trị vòng đời khách hàng (*Customer Lifetime Value*).

Cấu hình mô phỏng được thiết lập với thời gian dự báo $T = 6$ **tháng** và số lượng mô phỏng $N = 1000$ **kịch bản cho mỗi khách hàng**. Tại mỗi bước thời gian t, trạng thái của khách hàng i được cập nhật theo quy tắc:

$$S_t^{(i)} \sim \text{Categorical}\left(\mathbf{M}^{(i)}_{S_{t-1}^{(i)},\cdot}\right) \tag{10}$$

trong đó $\mathbf{M}^{(i)}_{S_{t-1}^{(i)},\cdot}$ là hàng tương ứng với trạng thái hiện tại trong ma trận chuyển trạng thái của khách hàng i.

V. Khung đánh giá

Để đánh giá hiệu quả của việc phân loại và xác suất của các mô hình dự đoán churn, nhóm sử dụng tập hợp các chỉ số đánh giá phổ biến trong phân loại nhị phân bao gồm Accuracy, Recall, AUC, F1-Score và Brier Score.

A. Accuracy (Độ chính xác)

Accuracy đo lường tỷ lệ dự đoán đúng trên tổng số mẫu quan sát, được xác định bởi:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{11}$$

trong đó TP, TN, FP và FN lần lượt biểu thị số lượng *true positives*, *true negatives*, *false positives* và *false negatives*. Mặc dù Accuracy cung cấp cái nhìn tổng quát, chỉ số này có thể gây hiểu lầm trong bối cảnh dữ liệu mất cân bằng.

B. Recall (Độ nhạy)

Recall đo lường khả năng mô hình phát hiện đúng các khách hàng thực sự rời bỏ (Churn), được tính như sau:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{12}$$

chỉ số này đặc biệt quan trọng trong bài toán churn do việc bỏ sót khách hàng có nguy cơ rời bỏ thường gây tổn thất lớn cho doanh nghiệp.

C. F1-score

F1-score là trung bình điều hòa giữa Precision và Recall, phản ánh sự cân bằng giữa khả năng phát hiện churn và độ chính xác của các dự đoán dương:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{13}$$

trong đó Precision $= \frac{\text{TP}}{\text{TP+FP}}$.

D. AUC (Area Under the ROC Curve)

AUC đánh giá khả năng mô hình xếp hạng đúng các mẫu dương và âm dựa trên điểm số xác suất đầu ra:

$$\text{AUC} = \Pr\left(s(x^+) > s(x^-)\right), \tag{14}$$

trong đó $s(\cdot)$ là hàm điểm số của mô hình, x^+ và x^- lần lượt là một mẫu thuộc lớp dương và lớp âm. Giá trị AUC càng cao thể hiện khả năng phân biệt càng tốt.

E. Brier Score

Brier Score đo lường chất lượng của các dự đoán xác suất bằng sai số bình phương trung bình giữa xác suất dự đoán và nhãn thực:

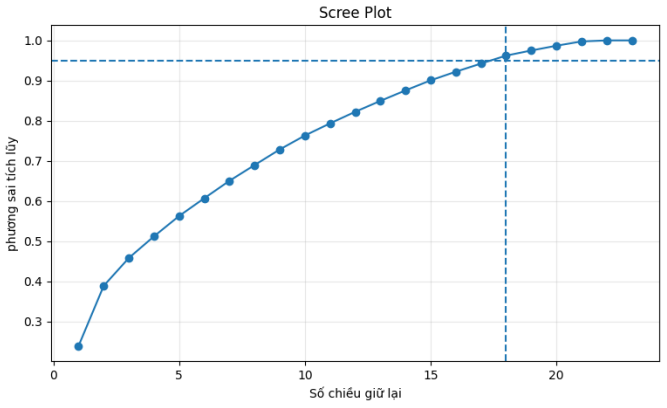
$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2, \tag{15}$$

trong đó p_i là xác suất dự đoán cho mẫu thứ i và $y_i \in \{0, 1\}$ là nhãn thực tương ứng. Giá trị Brier Score càng nhỏ cho thấy mô hình càng được hiệu chỉnh tốt về mặt xác suất.

VI. Kết quả và phân tích

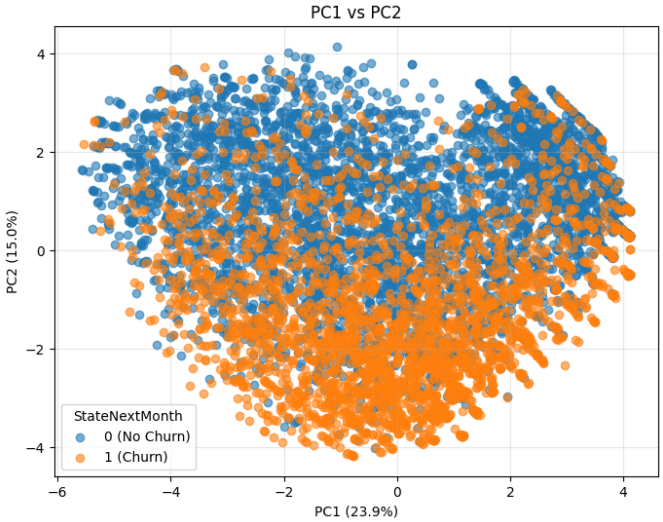
A. Phân tích giảm chiều dữ liệu

Hình 4 cho thấy phương sai tích lũy đạt **96,21%** tại **18 thành phần chính**. Sau ngưỡng này, mức cải thiện phương sai không đáng kể, xác nhận 18 thành phần là đủ để biểu diễn hiệu quả và bảo toàn thông tin quan trọng cho các thực nghiệm tiếp theo.



Hình 4. Phân tích phương sai tích lũy và trị riêng qua các thành phần chính (Scree Plot).

Hình 5 minh họa phân bố dữ liệu trên hai thành phần chính đầu tiên (PC1 và PC2), lần lượt giải thích 23.9% và 15.0% phương sai. Các mẫu thuộc lớp *Churn* và *No Churn* cho thấy sự chồng lấn đáng kể và hình thành các cụm cục bộ, trong khi ranh giới phân lớp trong không gian PC1–PC2 không mang tính tuyến tính rõ ràng. Điều này cho thấy các mô hình tuyến tính đơn giản khó đạt hiệu năng cao và củng cố tính phù hợp của các mô hình học máy phi tuyến trong bài toán dự đoán churn.



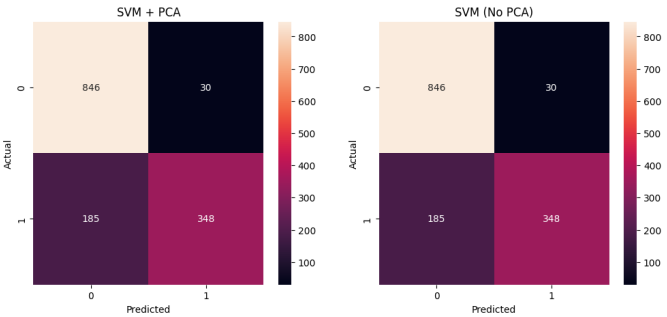
Hình 5. Biểu đồ phân bố dữ liệu trên thành phần PC1 và PC2.

B. Hiệu năng phân loại của mô hình

Kết quả so sánh hiệu năng giữa mô hình **SVM** truyền thống và mô hình **SVM** kết hợp **PCA** được chi tiết hóa trong Bảng I và Hình 6.

Bảng I
So sánh hiệu năng giữa mô hình SVM (PCA) và SVM (No PCA)

Chỉ số	SVM + PCA	SVM (No PCA)
Train time (s)	**7.406240**	8.769527
Accuracy	0.847410	0.847410
Recall	0.652908	0.652908
F1-score	0.763996	0.763996
AUC	**0.835495**	0.831484
Brier score	**0.1264**	0.1266



Hình 6. Ma trận nhầm lẫn (Confusion Matrix) so sánh kết quả dự báo của mô hình SVM và SVM-PCA.

Mặc dù không có sự khác biệt về các chỉ số phân loại truyền thống như **Accuracy**, **Recall** và **F1-score**, mô hình **SVM kết hợp PCA** đạt giá trị **Brier score thấp hơn** so với mô hình **SVM không giảm chiều** (0.1264 so với 0.1266).

Kết quả này cho thấy việc áp dụng **PCA** không làm suy giảm chất lượng dự báo xác suất, mà ngược lại còn **cải thiện nhẹ mức độ hiệu chỉnh xác suất đầu ra**. Điều này hàm ý rằng xác suất rời bỏ do mô hình **SVM–PCA** ước lượng phản ánh sát hơn tần suất xảy ra thực tế so với mô hình huấn luyện trên dữ liệu gốc.

Trong bối cảnh bài toán động học khách hàng, nơi xác suất dự báo được sử dụng làm đầu vào cho **ma trận chuyển trạng thái của mô hình Markov lai**, việc đạt **Brier score thấp hơn** là đặc biệt quan trọng. Nhờ đó, xác suất dự báo từ mô hình **SVM–PCA** có thể được sử dụng trực tiếp để xấp xỉ xác suất chuyển trạng thái:

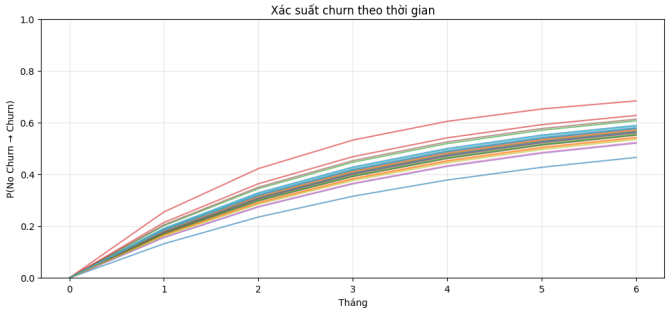
$$P(\text{Active} \to \text{Churn}) \approx p_{\text{SVM–PCA}}, \tag{16}$$

đảm bảo tính nhất quán và ổn định của mô hình Markov trong các bước mô phỏng tiếp theo.

C. Phân tích Xích Markov lai và Mô phỏng Monte Carlo

Trong thí nghiệm này, nhóm lựa chọn **30 khách hàng** có trạng thái ban đầu là *No Churn* nhằm phân tích động thái chuyển trạng thái theo thời gian. Với mỗi khách hàng, một **xích Markov lai** (*Hybrid Markov Chain*) được xây dựng, trong đó ma trận chuyển trạng thái có dạng:

$$\mathbf{P} = \begin{pmatrix} p_{\text{NoChurn}\to\text{NoChurn}} & p_{\text{NoChurn}\to\text{Churn}} \\ 0.0872 & 0.9128 \end{pmatrix} \tag{17}$$



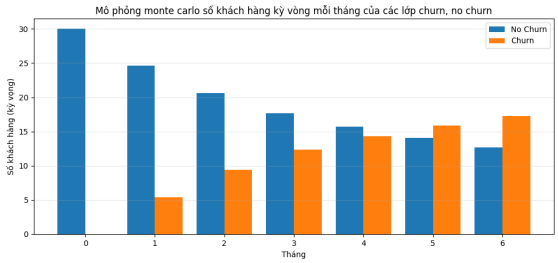
Hình 7. Kết quả mô phỏng Monte Carlo dự báo quỹ đạo xác suất rời bỏ cá nhân hóa theo thời gian.

Dù cùng trạng thái ban đầu, đường cong xác suất theo thời gian vẫn cho thấy sự phân tán lớn như hình 7, phản ánh rủi ro rời bỏ đặc thù của từng khách hàng mà mô hình *Markov* truyền thống không thể nắm bắt.

Hình 8 cho thấy kết quả mô phỏng Monte Carlo (1000 lần lặp) phù hợp với lý thuyết, phản ánh tính bất định trong hành vi khách hàng và chỉ ra rằng nếu không có can thiệp, xác suất rời bỏ tăng dần theo thời gian, đặc biệt là nhóm khách hàng có xác suất Churn cao ngay từ ban đầu.

VII. Kết luận

Bài báo đề xuất khung phương pháp PCA-SVM kết hợp xích Markov lai và mô phỏng Monte Carlo để dự báo rời bỏ. Kết quả cho thấy PCA giảm chiều hiệu quả mà vẫn bảo toàn



Hình 8. Kết quả mô phỏng Monte Carlo dự báo số lượng khách hàng theo thời gian.

độ chính xác, trong khi mô hình Markov và mô phỏng Monte Carlo giúp cá nhân hóa rủi ro theo thời gian, hỗ trợ doanh nghiệp ra quyết định chiến lược. Trong tương lai, nhóm dự kiến tích hợp học sâu (LSTM) và kỹ thuật XAI (SHAP) để nâng cao độ chính xác và tính minh bạch, đồng thời tối ưu hóa thuật toán cho dữ liệu lớn thời gian thực.

VIII. Phân công công việc

Bảng II
Bảng phân công công việc của nhóm 19

Thành viên	Nhiệm vụ chính	Hoàn thành (%)
Nguyễn Bảo Quân	Xây dựng và triển khai các đoạn mã báo cáo đồ án và thực hành; phát triển nội dung chính của đề tài	100
Trần Minh Phát	Hỗ trợ phát triển nội dung; biên soạn tài liệu LaTex và thiết kế bài trình bày PowerPoint.	100
Nguyễn Đình Tuấn Phúc	Hỗ trợ phát triển nội dung; biên soạn tài liệu Latex và thiết kế bài trình bày PowerPoint.	100

Tài liệu

[1] W. Verbeke, D. Martens, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach," *European Journal of Operational Research*, 2012.

[2] C.-P. Wei and I.-T. Chiu, "Turning telecommunications call details into churn prediction: a data mining approach," *Expert systems with applications*, vol. 23, no. 2, pp. 103–112, 2002.

[3] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data*, vol. 6, pp. 1–24, 2019.

[4] A. A. Ahmed and D. Maheswari, "Churn prediction in telecommunication for high-dimensional data using pca," *International Journal of Pure and Applied Mathematics*, 2017.

[5] O. Netzer, J. M. Lattin, and V. Srinivasan, "A discrete-time hidden markov model for customer relationship management," *Marketing Science*, vol. 27, no. 2, pp. 185–210, 2008.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[7] I. T. Jolliffe, *Principal Component Analysis*. Springer Science & Business Media, 2002.

[8] P. Brémaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer Science & Business Media, 2013.