ỨNG DỤNG CỦA PCA, MONTE CARLO VÀ NAIVE BAYES VÀO CÁC BÀI TOÁN THỰC TẾ

Trần Thái Hoà
Khoa Khoa học và Kỹ thuật thông tin
Trường Đại học Công nghệ thông tin
Thành phố Hồ Chí Minh, Việt Nam
21522082@gm.uit.edu.vn

Võ Hoàng An
Khoa Khoa học và Kỹ thuật thông tin
Trường Đại học Công nghệ thông tin
Thành phố Hồ Chí Minh, Việt Nam
21520555@gm.uit.edu.vn

Nguyễn Thị Huyền Trang
Khoa Khoa học và Kỹ thuật thông tin
Trường Đại học Công nghệ thông tin
Thành phố Hồ Chí Minh, Việt Nam
21520488@gm.uit.edu.vn

Tóm tắt – Trong đồ án môn học này, nhóm lần lượt giải quyết các bài toán nhỏ dựa trên các phương pháp thống kê và xác suất đã học.

Từ khoá – Gaussian Naive Bayes, Monte Carlo, Principal Component Analysis

A. MONTE CARLO SIMULATION

Tóm tắt – Nhóm dùng mô phỏng Monte Carlo từ bộ dữ liệu bán hàng của năm trước để tạo ra một bộ dữ liệu mới, sau đó nhóm tiến hành phân tích bộ dữ liệu mới đó để dự đoán tổng số tiền hoa hồng phải trả cho nhân viên trong năm tiếp theo.

I. Giới thiệu bài toán

Tiền hoa hồng là số tiền thù lao mà người ủy thác trả cho người trung gian (làm đại lí hay môi giới) về những dịch vụ đã làm tùy thuộc tính chất và khối lượng công việc. Việc dự đoán số tiền hoa hồng cần phải trả trong năm kế tiếp sẽ góp phần giúp các doanh nghiệp tính toán tốt các khoản kinh phí dự trù. Có rất nhiều mô hình có thể giải quyết được bài toán dự đoán này. Nhóm đã chọn mô phỏng Monte Carlo để giải quyết bài toán vì tính đơn giản cũng như độ hiệu quả cao.

Đầu vào của mô hình là giá trị trung bình và độ lệch chuẩn của biến tổng số tiền hoa hồng Commision Amount trong bộ dữ liệu của quá khứ. Đầu ra của mô hình là biến dự đoán Commision Amount_Pre là tổng số tiền hoa hồng ước lượng của năm tiếp theo.

II. Giới thiệu bộ dữ liệu

Bộ dữ liệu [1] quá khứ gồm các thuộc tính: Sales Rep, Sales Target, Actual Sales, Commision Amount, Percent to Plan, Commision Rate.

Trong đó:

Sales Rep là thứ tự của nhân viên bán hàng

Percent to Plan = (Actual Sales / Sales Target) tuân theo phân phối chuẩn với giá trị trung bình là 1 và độ lệch chuẩn là 0.1.

$$\textbf{CommisionRate} = \begin{cases} 0.02 \text{ nếu Percent to Plan } <= 0.9 \\ 0.03 \text{ nếu Percent to Plan } <= 0.99 \\ 0.04 \text{ trong các trường hợp còn lại} \end{cases}$$

Commision Amount = CommisionRate * Actual Sales.

Sales Target là doanh số bán hàng dự kiến, gồm 6 giá trị (75000, 100000, 200000, 300000, 400000, 500000)

Actual Sales là doanh số thực tế đạt được.

III. Phương pháp thực hiện

1. Định luật giới hạn trung tâm: [2]

Là một lí thuyết thống kê cho rằng với cỡ mẫu đủ lớn từ một tổng thể có phương sai hữu hạn, giá trị trung bình của tất cả các mẫu sẽ xấp xỉ bằng trung bình của tổng thể bất chấp hình dạng phân phối dữ liệu trong thực tế. Điều này cực kỳ hữu ích trong việc dự đoán các đặc điểm của tổng thể.

$$\lim_{n\to\infty} P\left(\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \leq z\right) = \phi(z)$$

Trong đó:

$$\overline{X_n} = \frac{S_n}{n} = \frac{X_1 + \cdots + X_n}{n}$$

2. Mô phỏng Monte Carlo [3]

Mô phỏng Monte Carlo là kỹ thuật để hiểu tác động của các rủi ro và tính không chắc chắn trong tài chính, quản lý dự án, chi phí và các mô hình dự báo.

Kỹ thuật Monte Carlo bao gồm ba bước cơ bản:

- Thiết lập mô hình dự đoán.
- Xác định phân phối xác suất của các biến độc lập.
- Chạy mô phỏng lặp đi lặp lại cho đến khi thu thập đủ kết quả để tạo thành một mẫu đại diện cho số lượng gần như vô hạn các kết hợp có thể có.

Phương pháp mô phỏng Monte Carlo là một phương pháp mô phỏng bằng xác suất. Phương pháp chủ yếu dựa trên hai luật quan trọng của xác suất là luật số lớn và luật số lớn yếu.

IV. Cài đặt thí nghiệm

Nhóm dùng thư viện numpy của python để tạo ra bộ dữ liệu mới, cụ thể như sau:

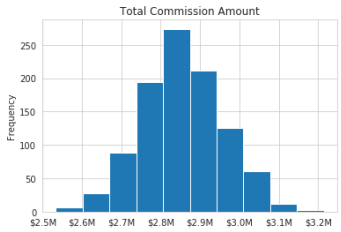
- Tạo dữ liệu mới của thuộc tính **Percent to Plan** với kỳ vọng là 1 và độ lệch chuẩn là 0.1.
- Tạo dữ liệu mói của thuộc tính **Sales Target** là các giá trị (75000, 100000, 200000, 300000, 400000, 500000) với xác suất tương ứng (0.3, 0.3, 0.2, 0.1, 0.05, 0.05).
- Tính toán hai thuộc tính trên để tạo ra dữ liệu của các thuộc tính **Actual Sales**, **Commision Rate** và **Commision Amount**.
- Chạy mô phỏng Monte Carlo với số lượng nhân viên bán hàng là 500 người và số lần chạy mô phỏng là 1000.

V. Kết quả thí nghiệm:

1. Kết quả

Mô phỏng Monte Carlo cho kết quả như sau:

	Sales	Commision_ Amount	Sales_ Target
Mean	83,617,936.0	2,854,916.1	83,619,700.0
Std	2,727,222.9	103,003.9	2,702,621.8



Tổng số tiền hoa hồng dự đoán có giá trị trung bình khoảng $2.85M và độ lệch chuẩn là $103K.

2. Phân tích lỗi

Mô phỏng Monte Carlo phụ thuộc rất nhiều vào các thuộc tính đầu vào và hàm phân phối xác suất của chúng. Nếu chúng ta không làm tốt trong khâu xác định này, thì mô hình phỏng đoán có thể cho ra kết quả không như mong đợi.

Việc cố gắng chạy càng nhiều mô phỏng không phải lúc nào cũng mang lại hiệu quả cao, bởi vì nó có thể làm giảm hiệu suất của các mô hình tính toán.

VI. Kết luận

Mô phỏng Monte Carlo cung cấp nhiều kết quả có thể xảy ra và xác suất của mỗi kết quả từ một tập lớn các mẫu dữ liệu ngẫu nhiên.

Mô phỏng Monte Carlo tương đối dễ hiểu và dễ sử dụng đối với những người không có nền tảng toán học sâu sắc.

B. GAUSSIAN NAIVE BAYES

Tóm tắt – Trong báo cáo này Nhóm xây dựng mô hình dự đoán khách hàng dựa trên các dữ liệu về giới tính, tuổi và mức lương. Đây là bài toán phân loại với 2 thuộc tính: Age và EstimatedSalary. Kết quả đem lại sẽ cho thấy độ hiệu quả của thuật toán Gaussian Naive Bayes trong việc xây dựng mô hình cho bài toán phân lớp khách hàng.

I. Giới thiệu bài toán

Phân loại khả năng mua hàng của khách hàng là công việc cần thiết để nắm bắt những đặc điểm chung trong dữ liệu khách hàng, thống kê những đặc tính tương đồng của những khách hàng của doanh nghiệp từ đó chia các đối tượng khách hàng thành những nhóm nhỏ để định hình chính xác nhóm khách hàng mà doanh nghiệp muốn hướng và đem đến cho họ một dịch vụ tốt nhất. Bên cạnh đó để doanh nghiệp có thể giảm bớt được chi phí cũng như thời gian phải phục vụ.

Đầu vào của bài toán là các thông số như: Tuổi, Lương của khách hàng. Đầu ra của bài toán cho biết Khách hàng đó có phải là khách hàng tiềm năng của doanh nghiệp đó hay không.

Đầu Vào	Đầu ra
30,79000	0
27, 137000	1

Trong phần II tập trung giới thiệu về tập dữ liệu có sẵn, phân tích sơ bộ tập dữ liệu. Phần III trình bày tóm tắt các thuật toán mà nhóm sử dụng để huấn luyện mô hình cho bài toán.. Phần IV trình bày ứng dụng cho bài toán và kết quả đạt được. Phần V phân tích lỗi và đưa ra hướng phát triển của nhóm trong tương lai.

II. Giới thiệu bộ dữ liệu

Bộ dữ liệu Social_Network_Ads [4] (đã được chuẩn hoá) bao gồm 5 cột: UserID, Gender, EstimatedSalary, Age, Purchased, gồm 400 dòng dữ liệu.

Nhóm chỉ sử dụng 3 thuộc tính là Age, EstimatedSalary và Purchased. Trong đó thuộc tính Purchased là thuộc tính mục tiêu cho mô hình bài toán.

Thuộc tính Purchased gồm 2 giá trị :1 – Mua, 0- Không mua.

Nhóm tiến hành chia tập dữ liệu thành 2 tập, huấn luyện (train set) và kiểm thử (test set) để thí nghiệm mô hình. Hai tập dữ liệu được chia ngẫu nhiên theo tỉ lệ train:test là 8:2.

Tập dữ liệu	Số lượng
Train	320
Test	80

III. Phương pháp thực hiện

1. *Mô hình phân lớp*

Là một mô hình machine learning dùng để phân loại các vậy mẫu đưa trên các thuộc tính đã xác định.

2. *Định lý Bayes* [5]

Định lý Bayes [5] tìm xác suất của một sự kiện xảy ra với xác suất của một sự kiện khác đã xảy ra.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

3. *Naive Bayes* [6]

Naive Bayes [6] là một nhóm các thuật toán phân loại học máy có giám sát dựa trên định lý Bayes. Đây là một kỹ thuật phân loại đơn giản, nhưng có hiệu quả cao trong thực tế. Naive Bayes giả định mọi cặp thuộc tính đều độc lập với nhau.

$$\mathrm{P}(\mathrm{y}|x_1, \dots, x_n) = \frac{\left(P(x_1|y)P(x_2|y)\dots P(x_n|y)\right)}{\left(P(x_1)P(x_2)\dots P(x_n)\right)}$$

Có thể được diễn đạt như sau:

$$\mathrm{P}(\mathrm{y}|x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{\left(P(x_1)P(x_2)\dots P(x_n)\right)}$$

Vì mẫu số không đổi đối với một đầu vào nhất định, chúng ta có thể loại bỏ thuật ngữ đó:

$$\mathrm{P}(\mathrm{y}|x_1, \dots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

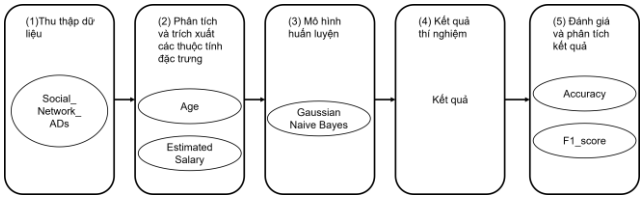
4. *Gaussian Naive Bayes* [5]

Gaussian Naive Bayes [5] nhận các thuộc tính có giá trị liên tục và tuân theo phân phối chuẩn.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma^2}}$$

IV. Cài đặt thí nghiệm

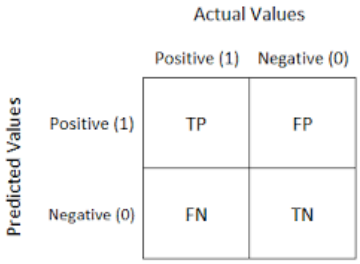
1. Quy trình thí nghiệm



2. Độ đo đánh giá

Nhóm sử dụng 2 độ đo là Accuracy và F1-score-macro để đánh giá mô hình bài toán.

- **Confusion Matrix** [7]**:** là một ma trận vuông với kích thước mỗi chiều bằng số lượng lớp dữ liệu. Giá trị tại hàng thứ i, cột thứ j là số lượng điểm lẽ ra thuộc vào class i nhưng lại được dự đoán là thuộc vào class j.



- **Accuracy** [7]: là tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử.

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

- **F1-score** [7]: là một trung bình điều hoà (harmonic mean) của precision và recall. Nó có xu hướng lấy giá trị gần với giá trị nào nhỏ hơn giữa 2 giá trị Precision và Recall và đồng thời nó có giá trị lớn nếu cả 2 giá trị Precision và Recall đều lớn.

$$F1 - score == \frac{2 * precision * recall}{precision + recall}$$

- **Precision** là tỉ lệ số điểm true positive trong số những điểm được phân loại là positive (TP + FP).
- **Recall** là tỉ lệ số điểm true positive trong số những điểm thực sự là positive (TP + FN).

V. Kết quả thí nghiệm

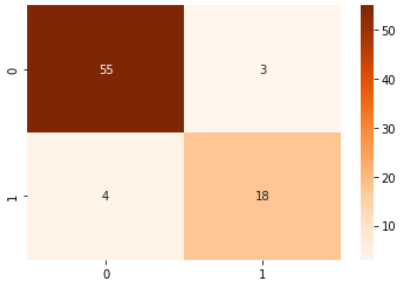
1. Kết quả

Qua đó với bài toán này và tập dữ liệu của nhóm, mô hình Gaussian Naive Bayes cho kết quả như sau:

Accuracy	F1-score
91.25%	88.87%

Cho thấy mô hình GaussianNB dự đoán với kết quả cao đối với bộ dữ liệu Social_Network_Ads.

`Đây là confusion matrix của mô hình:



2. Phân tích lỗi

Tuy nhiên có nhược điểm lớn là yêu cầu các đặc trưng đầu vào phải độc lập, mà điều này khó xảy ra trong thực tế làm giảm chất lượng của mô hình.

VI. Kết luận

1. Kết luận

Qua bài toán này chúng tôi đã cài đặt thành công đối với mô hình Gaussian Naïve Bayes cho bài toán phân loại khả năng mua hàng của khách hàng cho kết quả đạt 91.25% đối với độ đo accuracy và 88.87% đối với độ đo F1-score. Vì vậy, nhóm kết luận rằng 2 thuộc tính Age và EstimatedSalary có tầm ảnh hưởng đến khả năng mua hàng của khách hàng.

2. Hướng phát triển

Trong tương lai nhóm sẽ mở rộng bộ dataset, thử nghiệm các phương pháp tiền xử lý mới và thử nghiệm trên các mô hình học sâu kết hợp tinh chỉnh các mô hình.

C. PRINCIPAL COMPONENT ANALYSIS

Tóm tắt – Trong bài toán này, nhóm xây dựng mô hình dự đoán các yếu tố ảnh hưởng nhiều đến tỷ lệ khách hàng rời đi [8] **dựa trên tập hợp các yếu tố của bộ dữ liệu mà nhóm thu thập được. Bài toán sử dụng thuật toán Principal Component Analysis (PCA)** [8] **để giảm số chiều của dữ liệu trước khi tiến hành dự đoán trên mô hình hồi quy logistic. Kết quả đem lại sẽ cho thấy các yếu tố ảnh hưởng đến tỷ lệ khách hàng rời đi cũng như độ hiệu quả của thuật toán PCA** [8] **trong việc xây dựng mô hình cho bài toán này.**

I. Giới thiệu bài toán

Tỷ lệ khách hàng rời đi [9] là tỷ lệ dựa trên lượng khách hàng rời đi và khách hàng hiện có trong một khoảng thời gian nhất định. Đây là một trong những chỉ số quan trọng hàng đầu của công ty, doanh nghiệp. Nếu lượng khách hàng hủy hợp đồng hay ngưng sử dụng dịch vụ tăng cao, thì đây chính là dấu hiệu của việc kinh doanh có vấn đề. Do đó, việc dự đoán các yếu tố có ảnh hưởng đến tỷ lệ này sẽ giúp các doanh nghiệp tìm ra các giải pháp kịp thời để giữ chân khách hàng.

Đầu vào của bài toán là các thông tin (thuộc tính) của khách hàng như: độ tuổi, giới tính, số lượng giao dịch, ….

Đầu ra của bài toán là các thuộc tính có khả năng cao ảnh hưởng đến tỷ lệ khách hàng rời đi

II. Giới thiệu bộ dữ liệu

Bộ dữ liệu BankChurners (đã được chuẩn hoá) gồm 10,000 dòng với 18 thuộc tính là các thông tin của khách hàng (Customer_Age, Gender, Dependent_count, …) nhưng nhóm chỉ giữ lại 11 thuộc tính. Trong đó thuộc tính 'Attrition_Flag' (gồm 2 giá trị là 'Existing customer' và 'Attrited customer' là thuộc tính mục tiêu của mô hình bài toán.

Tập dữ liệu được chia thành hai phần là tập huấn luyện (train set) và tập kiểm thử (test set) để thí nghiệm mô hình, 2 tập dữ liệu được chia ngẫu nhiên theo tỉ lệ train:test là 7:3. Cụ thể được thể hiện ở bảng sau:

Tập dữ liệu	Số lượng
Train	7000
Test	3000

III. Phương pháp thực hiện

1. Principal Components Analysis (PCA) [8]

PCA [8]là một phương pháp biến đổi giúp giảm số lượng lớn các biến có tương quan với nhau thành tập ít các biến sao cho các biến mới tạo thành (Principal Components - PCs) là tổ hợp tuyến tính của những biến cũ, không có tương quan tuyến tính với nhau mà vẫn giữ được nhiều nhất lượng thông tin từ nhóm biến ban đầu.

Giả sử chúng ta có một vector ngẫu nhiên X

$$\mathrm{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

Có ma trận hiệp phương sai là

$$var(X) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}$$

Các biến Y_i (PC_i) có được từ sự kết hợp tuyến tính các biến lại với nhau:

$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

Phương sai và hiệp phương sai của các Y_i sẽ được tính như sau:

$$var(Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{il}\sigma_{kl} = \mathbf{e}_i'\Sigma\mathbf{e}_i$$

$$\mathrm{cov}(Y_i, Y_j) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{jl}\sigma_{kl} = \mathbf{e}_i'\Sigma\mathbf{e}_j$$

Các hệ số e_i hay còn gọi là trọng số thể hiện mức độ ảnh hưởng của các biến cũ (X_i) vào biến mới (Y_i)

PCA [8] hoạt động bằng cách tìm các hệ số e_i sao cho var(Y_i) là lớn nhất thông qua các giá trị riêng và vector riêng của các ma trận hiệp phương sai của các biến X_i.

Đối với Y_1 (PC_1), nó phải thỏa điều kiện (1) và (2):

$var(Y_1) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{1k}e_{1l}\sigma_{kl} = \mathbf{e}_1'\Sigma\mathbf{e}_1$ lớn nhất (1)

$$\mathbf{e}_1'\mathbf{e}_1 = \sum_{j=1}^{p} e_{1j}^2 = 1 \text{ (2)}$$

Đối với các Y_i còn lại thì ngoài hai điều kiện trên còn phải thỏa thêm 1 điều kiện:

$$\mathrm{cov}(Y_1, Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{1k}e_{il}\sigma_{kl} = \mathbf{e}_1'\Sigma\mathbf{e}_i = 0$$

$$\mathrm{cov}(Y_2, Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{2k}e_{il}\sigma_{kl} = \mathbf{e}_2'\Sigma\mathbf{e}_i = 0$$

…

$$\mathrm{cov}(Y_{i-1}, Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{i-1,k}e_{il}\sigma_{kl} = \mathbf{e}_{i-1}'\Sigma\mathbf{e}_i = 0$$

2. Hồi quy logistic

Phân tích hồi quy logistic [8] là một kỹ thuật thống kê để xem xét mối liên hệ giữa biến độc lập (biến số hoặc biến phân loại) với biến phụ thuộc là biến nhị phân. Trong hồi quy logistic, biến phụ thuộc chỉ có hai trạng thái 1 và 0.

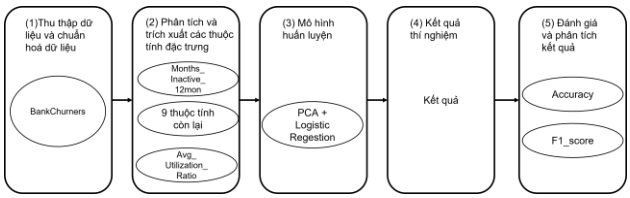
Phương trình hồi quy logistic có thể được xây dựng với các giá trị 0 và 1 được thu nhận từ kết quả đánh giá thuộc tính mục tiêu của khách hàng.

$$\frac{F(x)}{1-F(x)} = e^{\beta_0+\beta_1 x}$$

Trong đó, đầu vào là giá trị $\beta_0 + \beta_1 x$ và đầu ra là $F(x)$. Trong phân tích hàm nhiều biến, $\beta_0 + \beta_1 x$ có thể được sửa đổi thành $\beta_0 + \beta_1 x + \beta_2 x \dots + \beta_m xm$. Sau đó, khi được sử dụng trong các phương trình liên quan đến tỷ số odds với giá trị của các yếu tố dự báo, phương trình hồi quy tuyến tính sẽ trở thành hồi quy không tuyến tính với m biến, các thông số β_j cho tất cả j được ước tính.

IV. Cài đặt thí nghiệm

1. Quy trình thí nghiệm



2. Độ đo

Nhóm sử dụng độ đo Accuracy và F1-score đã được trình bày ở trên.

3. Xây dựng bộ dữ liệu mới

Sau khi áp dụng PCA [8], nhóm thu được bộ dữ liệu mới đơn giản hơn rất nhiều so với bộ dữ liệu cũ là:

	PC1	PC2	Attrition_Flag
0	0.28	-0.62	1
1	-0.61	1.43	1
…	…	…	…

Ngoài dùng hai PCs, nhóm cũng thử dùng đến 3 PCs và 4 PCs và cả bộ dữ liệu cũ cho mục đích so sánh độ chính xác của mô hình tính toán dựa trên thang đo accuracy và F1-score.

V. Kết quả:

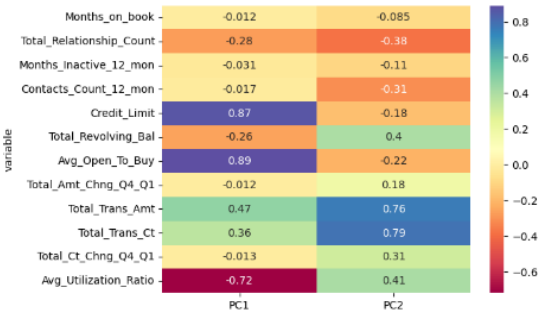
1. Độ Accuracy và F1 score

Số lượng PCs	Accuracy (%)	F1-score
Không dùng PCA	89	0.94
2	84	0.91
3	88	0.93
4	88	0.93

Với 3 PCs mô hình sẽ cho ra kết quả dự đoán tốt nhất, tuy nhiên chỉ với 2 PCs cũng cho ra được kết quả khá cao (84%).

2. Dự đoán các yếu tố ảnh hưởng đến tỷ lệ khách hàng rời đi

Vì 2PCs cho kết quả khá tốt, ta có thể dùng biểu đồ nhiệt (heatmap) để xem trọng số của các thuộc tính trong 2 PCs đó.



VI. Kết luận và hướng phát triển

1. **Kết luận:** Dựa vào biểu đồ 1, nhóm kết luận rằng có ba thuộc tính có khả năng cao ảnh hưởng đến tỷ lệ khách hàng rời đi, đó là: Total_Trans_Amt, Avg_Utilizatoin_Ratio và Total_Trans_Ct, điều này cũng hợp lí nếu ta xét đến các yếu tố có sức ảnh hưởng cao ở ngoài thực tế.
2. **Phân tích lỗi:** PCA vẫn còn hạn chế trong việc diễn đạt ý nghĩa của các biến mới (các PCs), cho nên sự kết luận của nhóm chưa thật sự tốt, vì có thể còn những thuộc tính khác cũng có sự ảnh hưởng cao đến tỷ lệ khách hàng rời đi mà nhóm đã không phát hiện ra.
3. **Hướng phát triển**: Trong tương lai nhóm có thể dùng phương pháp Varimax Rotation để tăng tính diễn đạt ý nghĩa của các PCs, cũng như sẽ có thêm các thang đo khác để xác định tối ưu hơn số `PCs cần dùng.

D. TÀI LIỆU THAM KHẢO

[1] C. Moffitt, "pbpython," 18 February 2019. [Online]. Available: https://pbpython.com/monte-carlo.html?fbclid=IwAR0PdLiGX9jQZq0O87aOgPL7WDTeqVG3IFaPVssrw7nPdCmTgAJ0P6LWXD0.

[2] "VietNamBiz," 7 November 2019. [Online]. Available: https://vietnambiz.vn/dinh-luat-gioi-han-trung-tam-central-limit-theorem-clt-la-gi-clt-trong-tai-chinh-20191105171517389.htm?fbclid=IwAR0kQtSzQWGNnHbvmpW1FB1IVlskFOYQIzdac25nUlxX5yEtv0gH86XAft4#:~:text=Định%20luật%20giới%20hạn%20trung%20tâm%20.

[3] AWS, "aws," [Online]. Available: https://aws.amazon.com/vi/what-is/monte-carlo-simulation/?fbclid=IwAR2ABYnihyJYHp7-r58w125NiOq9WW1bwC9EK4prAS-kyNZWilngpyxLgWQ.

[4] R. RAUSHAN, "kaggle," [Online]. Available: https://www.kaggle.com/datasets/rakeshrau/social-network-ads?resource=download.

[5] D. Xuân, 29 December 2020. [Online]. Available: https://cafedev.vn/tu-hoc-ml-bo-phan-loai-naive-bayes/.

[6] P. Majumder, "OpenGenusIQ," [Online]. Available: https://iq.opengenus.org/gaussian-naive-bayes/.

[7] "Bài 33: Các phương pháp đánh giá một hệ thống phân lớp," 3 January 2018. [Online]. Available: https://machinelearningcoban.com/2017/08/31/evaluation/.

[8] avcontentteam, 26 June 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/?fbclid=IwAR2pAuhOalFRDLEXmF41GdyioWjHWTXA5rW151CWrYgfAkVuPOZQC7l991Q.

[9] avcontentteam, 6 February 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/?fbclid=IwAR3XYJ0o1IGn0Ns3w0OOEHmbclD2pkEseiBl1hL7t1kxzaFLHRw7ytdUSbY.

[10] [Online]. Available: https://medium.com/m/global-identity?redirectUrl=https%3A%2F%2Ftowardsdatascience.com%2Fprincipal-component-analysis-with-python-an-example-for-beginners-by-a-beginner-ac052eff45c%3Ffbclid%3DIwAR0HwRzvajpShNLdQnniYtOeaEZtBhO2wZ9r1nuPNgSzlmG6aZyDWqTe4J0.

E. PHÂN CÔNG CÔNG VIỆC

Trần Thái Hoà	- Phân công công việc - Viết báo cáo - Thuyết trình (Bayes)
Võ Hoàng An	- Viết báo cáo - Làm Slide - Thuyết trình (PCA)
Nguyễn Thị Huyền Trang	- Viết báo cáo - Làm Slide - Thuyết trình (Monte)