1. Giới thiệu về DataSet

Dữ liệu liên quan đến các chiến dịch tiếp thị trực tiếp của một tổ chức ngân hàng Bồ Đào Nha. Các chiến dịch tiếp thị dựa trên các cuộc gọi điện thoại đến với khách hàng. Thường thì cần phải liên hệ với nhiều hơn một khách hàng để biết được sản phẩm (tiền gửi có kỳ hạn của ngân hàng) có được đăng ký ('có') hay không ('không').

Quy mô dữ liệu

- Số lượng dòng dữ liệu: 41.188
- Số lượng thuộc tính: 21

Dữ liệu khách hàng ngân hàng:

(1) Tuổi (age): Độ tuổi của khách hàng (kiểu số).
(2) Nghề nghiệp (job): Loại công việc của khách hàng (dữ liệu phân loại). Gồm các nhóm:
    - 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'
(3) Tình trạng hôn nhân (marital):
    - 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed.
(4) Trình độ học vấn (education):
    - 'Basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'
(5) Nợ quá hạn (default): Khách hàng có khoản nợ quá hạn không? ('no', 'yes', 'unknown').
(6) Vay thế chấp (housing): Khách hàng có khoản vay mua nhà không? ('no', 'yes', 'unknown').
(7) Vay cá nhân (loan): Khách hàng có khoản vay cá nhân không? ('no', 'yes', 'unknown').

Thông tin về lần liên lạc cuối cùng trong chiến dịch hiện tại

(8) Phương thức liên lạc (contact): 'cellular','telephone'
(9) Tháng liên lạc cuối cùng (month): Tháng trong năm từ tháng 1 đến tháng 12 ('jan', 'feb', 'mar',…, 'nov', 'dec')
(10) Ngày trong tuần của lần liên lạc cuối (day_of_week): Từ thứ Hai đến thứ Sáu ('mon','tue','wed','thu','fri')
(11) Thời lượng cuộc gọi (duration): Được tính bằng giây (kiểu số)

- Lưu ý quan trọng: Đây là một yếu tố ảnh hưởng mạnh đến kết quả dự đoán. Nếu duration = 0, chắc chắn khách hàng sẽ từ chối (y = no). Tuy nhiên, do thời lượng chỉ được biết sau khi cuộc gọi kết thúc, nên nếu sử dụng cho mô hình dự đoán thực tế, cần loại bỏ biến này để tránh rò rỉ dữ liệu

Thông tin khác về chiến dịch tiếp thị

(12) Số lần liên lạc trong chiến dịch (campaign): Tổng số lần liên hệ với khách hàng trong chiến dịch này (bao gồm cả lần liên hệ cuối)
(13) Số ngày kể từ lần liên hệ trước đó (pdays):Nếu khách hàng chưa từng được liên hệ trước đây, giá trị này sẽ là 999
(14) Số lần liên hệ trước chiến dịch hiện tại (previous): Số lần khách hàng đã được liên hệ trong các chiến dịch trước
(15) Kết quả của chiến dịch trước đó (poutcome): Có thể là 'failure', 'success' hoặc 'nonexistent (khách hàng chưa từng tham gia chiến dịch nào trước đó)

Thông tin về bối cảnh kinh tế - xã hội

(16) Tỷ lệ biến động việc làm (emp.var.rate): Được đo theo từng quý (dữ liệu số).
(17) Chỉ số giá tiêu dùng (cons.price.idx): Được đo hàng tháng (dữ liệu số).
(18) Chỉ số niềm tin người tiêu dùng (cons.conf.idx): Được đo hàng tháng (dữ liệu số).
(19) Lãi suất Euribor 3 tháng (euribor3m): Lãi suất trung bình của ngân hàng châu Âu trong 3 tháng gần nhất (dữ liệu số).
(20) Số lượng nhân viên trong nền kinh tế (nr.employed): Được đo theo từng quý (dữ liệu số).

Biến đầu ra

(21) y - khách hàng đã đăng ký gửi tiền có kỳ hạn chưa? ('yes', 'no')

2. Tiền xử lý

Quá trình tiền xử lý dữ liệu là một bước quan trọng trong bất kỳ bài toán học máy nào. Dữ liệu cần phải được chuẩn bị và xử lý một cách hợp lý để đảm bảo mô hình học máy có thể học được từ dữ liệu và đưa ra các dự đoán chính xác. Dưới đây là một mô tả chi tiết về các bước tiền xử lý mà nhóm nghiên cứu đã thực hiện trên bộ dữ liệu:

2.1. Kiểm tra và xử lý dữ liệu thiếu

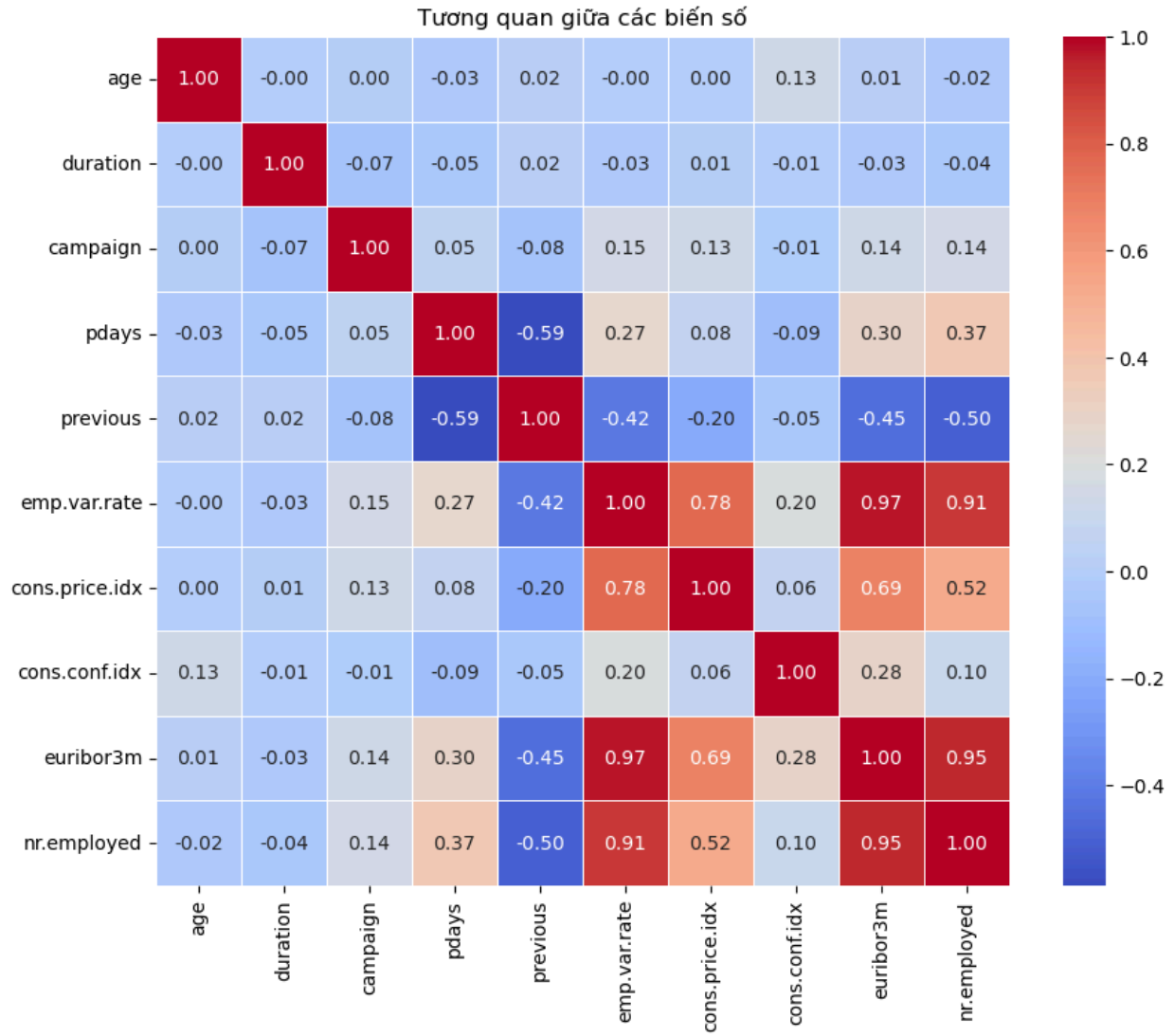
Dữ liệu ban đầu có thể chứa các giá trị bị thiếu, và điều quan trọng là phải xử lý chúng. Nhóm nghiên cứu đã sử dụng df.isnull().sum() để kiểm tra số lượng giá trị thiếu trong mỗi cột của DataFrame. Nhóm nhận ra bộ dữ liệu không chứa giá trị bị NULL.

2.2. Loại bỏ các giá trị trùng lặp

Dữ liệu trùng lặp có thể làm méo mó kết quả học máy. Nhóm nhận ra bộ dữ liệu có 12 dòng bị trùng lặp. Vì vậy, nhóm nghiên cứu đã loại bỏ các bản ghi trùng lặp bằng phương thức drop_duplicates(). Trong trường hợp này, nhóm đã giữ lại bản ghi đầu tiên của mỗi nhóm trùng lặp (keep='first') để đảm bảo dữ liệu không bị lặp lại.

2.3. Kiểm tra tương quan và giảm biến

Sau khi loại bỏ các giá trị trùng lặp, nhóm nghiên cứu tiếp tục tính toán ma trận tương quan giữa các biến số trong dữ liệu (các cột kiểu số). Việc này giúp nhận diện các biến có tương quan cao với nhau, từ đó hỗ trợ việc loại bỏ các cột thừa hoặc có sự trùng lặp thông tin giúp mô hình có thể hoạt động tốt hơn về cả mặt hiệu quả và thời gian.



Biểu đồ 2.1: Tương quan giữa các biến

Để trực quan hóa ma trận tương quan, nhóm nghiên cứu đã sử dụng seaborn để vẽ heatmap. Biểu đồ này giúp nhận diện mối quan hệ giữa các biến và đưa ra các quyết định loại bỏ những cột có tương quan quá cao, như trong ví dụ này là loại bỏ các cột 'emp.var.rate', 'euribor3m', 'nr.employed'.

2.4. Mã hóa các biến phân loại (Categorical Variables)

Các biến phân loại như job, marital, default, housing, loan, poutcome, và contact là các biến không có thứ tự, do đó, nhóm quyết định chuyển đổi các biến này thành các biến giả (dummy variables) bằng cách sử dụng phương pháp One-Hot Encoding thông qua pd.get_dummies(). Điều này giúp mô hình học máy có thể hiểu và xử lý các biến phân loại dưới dạng các cột số, thay vì dưới dạng văn bản.

Các biến có thứ tự như education, month, và day_of_week đã được ánh xạ thành các giá trị số:

- Biến education có các giá trị như 'illiterate', 'basic.4y', 'high.school', ... được ánh xạ thành các số nguyên từ 0 đến 7 vì giá trị của chúng có tính thứ tự.
- Biến month và day_of_week cũng được ánh xạ từ tên tháng và tên ngày thành các số nguyên tương ứng (ví dụ: 'jan' → 1, 'feb' → 2, ...; 'mon' → 1, 'tue' → 2, ...).

Biến mục tiêu y có giá trị là 'yes' hoặc 'no' (biến nhị phân). Để mô hình có thể xử lý được, nhóm nghiên cứu đã sử dụng LabelEncoder để chuyển đổi các giá trị này thành các giá trị số 0 và 1. Cụ thể, 'no' được mã hóa thành 0 và 'yes' thành 1.

3. Phân cụm khách hàng

Nhóm đã sử dụng thuật toán KMean để tiến hành nhóm khách hàng theo đặc trưng nhân khẩu học và tài chính. Quá trình phân cụm khách hàng bằng thuật toán K-Means được thực hiện qua các bước sau:

Trước tiên, nhóm loại bỏ một số biến không cần thiết như y, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed, month, day_of_week, duration, campaign, pdays và previous được loại bỏ khỏi tập dữ liệu ban đầu vì do các biến đó không thuộc nhóm đặc trưng nhân khẩu học và tài chính của khách hàng. Sau đó, PCA (Phân tích thành phần chính) được sử dụng để giảm chiều dữ liệu, giúp loại bỏ các đặc trưng dư thừa và tối ưu hóa quá trình phân cụm. Số lượng thành phần chính tối ưu được xác định dựa trên tiêu chí Kaiser (eigenvalue ≥ 1) và ra được con số 3 thành phần. Sau đó, dữ liệu được chuyển đổi thành không gian mới với số chiều giảm xuống, giúp cải thiện hiệu suất của thuật toán K-Means.

Sau khi có dữ liệu thì nhóm sẽ tiến hành hai phương pháp là Phương pháp Elbow và Phương pháp Silhouette Score để xác định số cụm tối ưu (k). Kết quả tại cả hai

phương pháp đều cho thấy số lượng cụm tối ưu là bằng 2 và số điểm Silhouette Score là 0.5859 cho thấy sự khác biệt trong các cụm.

K_means_segments	age	education	job_blue-collar	job_entrepreneur	job_housemaid	job_management	job_retired	job_self-employed	job_services	job_student
0	33.503591	4.443277	0.224905	0.031169	0.017609	0.059206	0.001833	0.034645	0.105118	0.033231
1	51.406775	3.995065	0.224393	0.042678	0.039944	0.091624	0.111363	0.034276	0.081022	0.000333

job_technician	job_unemployed	job_unknown	marital_married	marital_single	marital_unknown	default_unknown	default_yes	housing_unknown	housing_yes
0.182238	0.024064	0.004584	0.525248	0.392628	0.001910	0.155806	0.000038	0.023759	0.524599
0.131235	0.025607	0.014004	0.744865	0.085690	0.002001	0.301214	0.000133	0.024540	0.522606

loan_unknown	loan_yes	poutcome_nonexistent	poutcome_success	contact_telephone
0.023759	0.151986	0.861421	0.031780	0.362338
0.024540	0.151307	0.866831	0.036076	0.370432

Hình 3.1: Bảng thống kê mô tả các giá trị trung bình trong các cụm

Trên hình 3.1 là bảng thống kê mô tả giá trị trung bình của các cụm. Dựa trên kết quả này, nhóm đã rút ra được những nhận xét như về các cụm khách hàng như sau:

- Nhóm 0: Khách hàng có xu hướng vay mua nhà cao (housing_yes ~ 57%)

- Đặc trưng nhân khẩu học: Đây là nhóm có độ tuổi trung bình khoảng 40.1 tuổi. Trình độ học vấn trung bình cao hơn nhóm 1 (4.48 so với 4.08). Tỉ lệ khách hàng đã kết hôn cao (57.7%), thấp hơn nhóm 1 nhưng vẫn chiếm đa số.

- Hành vi tài chính và vay vốn: Tỷ lệ khách hàng có khoản vay mua nhà cao (56.7%), nhưng tỷ lệ vay tiêu dùng thấp (15.3%). Tỷ lệ vỡ nợ (default_yes) rất thấp (0.01%), cho thấy đây là nhóm khách hàng có độ tin cậy cao về tài chính. Một điểm đặc biệt chính là phần lớn chưa từng tham gia các chương trình tiếp thị trước đây (poutcome_nonexistent: 72.7%)

- Hành vi tiềm năng: Nhóm này có xu hướng ưu tiên sở hữu nhà, do đó có thể quan tâm đến các gói vay thế chấp hoặc tái cấp vốn. Việc phần lớn chưa từng tham gia cho thấy cần

khai thác thêm bằng cách tiếp cận với các chiến dịch tiếp thị trực tiếp (như qua điện thoại hoặc email) để tăng tỷ lệ tham gia.

- Nhóm 1: Khách hàng có xu hướng phản hồi cao với chiến dịch tiếp thị (poutcome_success = 100%)

- Đặc trưng nhân khẩu học: Tuổi trung bình là tầm 39.9 tuổi. Trình độ học vấn thấp hơn một chút so với nhóm 0 (4.08 so với 4.48). Tỷ lệ khách hàng đã kết hôn cao hơn (63.3%).

- Hành vi tài chính và vay vốn: Tỷ lệ vay mua nhà thấp hơn nhóm 0 (48%), nhưng tỷ lệ thành công trong các chiến dịch tiếp thị rất cao (100%). Không có khách hàng nào bị mặc định (default_yes = 0%), cho thấy đây cũng là một nhóm khách hàng đáng tin cậy về tài chính. Tỷ lệ sử dụng phương thức liên hệ qua điện thoại cố định cao (contact_telephone = 63.9%).

- Hành vi tiềm năng: Nhóm này phản hồi tốt với các chiến dịch tiếp thị trước đây, vì vậy có thể tiếp tục áp dụng các chiến dịch marketing tương tự để thu hút họ. Phương thức liên hệ hiệu quả nhất với nhóm này là qua điện thoại, do đó nên ưu tiên các chiến dịch gọi điện trực tiếp.

4. Phân nhóm khách hàng

4.1. Quy Trình Xây Dựng và Đánh Giá Mô Hình

Trong quá trình nghiên cứu, nhóm đã áp dụng một quy trình rõ ràng và có hệ thống để xây dựng và huấn luyện các mô hình học máy, bao gồm Logistic Regression, Random Forest và K-Nearest Neighbors (KNN). Các bước thực hiện bao gồm chia dữ liệu, áp dụng **k-fold cross-validation**, tối ưu tham số bằng **RandomizedSearchCV**, và cuối cùng đánh giá mô hình.

4.1.1. Chia Dữ Liệu (Split 8 - 2)

Đầu tiên, nhóm chia dữ liệu thành hai phần: **80% dữ liệu được sử dụng để huấn luyện mô hình** và **20% còn lại được sử dụng để kiểm tra mô hình**. Quy trình chia dữ liệu theo tỷ lệ 80/20 giúp mô hình học từ một tập huấn luyện lớn trong khi vẫn giữ được một tập kiểm tra đủ lớn để đánh giá hiệu suất của mô hình một cách chính xác và khách quan.

4.1.2. K-Fold Cross-Validation

Nhóm áp dụng phương pháp **K-fold cross-validation** (với k = 5) để đánh giá mô hình trên 80% dữ liệu huấn luyện. Dữ liệu huấn luyện được chia thành 5 phần, mỗi lần một phần làm tập kiểm tra và 4 phần còn lại làm tập huấn luyện. Quá trình này giúp giảm

thiểu overfitting và cung cấp cái nhìn chính xác hơn về hiệu suất mô hình trên dữ liệu chưa thấy.

4.1.3. Tinh chỉnh tham số với RandomizedSearchCV

Sau khi chia dữ liệu và áp dụng k-fold cross-validation, nhóm sử dụng **RandomizedSearchCV** để tìm kiếm các tham số tối ưu cho ba mô hình học máy: **Logistic Regression**, **Random Forest**, và **K-Nearest Neighbors (KNN)**. RandomizedSearchCV thực hiện tìm kiếm tham số ngẫu nhiên trong một không gian tham số lớn, giúp tiết kiệm thời gian và tài nguyên so với việc kiểm tra tất cả các kết hợp có thể có, như trong GridSearchCV.

4.2. Xây dựng mô hình:

Dựa trên bộ dữ liệu đã được làm sạch, nhóm tiến hành sử dụng các thuật toán phân loại bao gồm: KNN, Random Forest, Hồi quy Logistic. Quy trình thực hiện bài toán này diễn ra như sau:

Sau khi train mô hình, nhóm dùng mô hình đó đánh giá cho dữ liệu test (20% bộ dữ liệu ban đầu) để đánh giá kết quả cho từng mô hình. Đối với bài toán của nhóm, nhóm cho rằng chỉ số **Recall** là thước đo quan trọng nhất bởi nhóm không muốn bỏ sót các khách hàng thực sự đồng ý sử dụng dịch vụ. Vì vậy nhóm sẽ sử dụng chỉ số này để lựa chọn mô hình tốt nhất.

	precision	recall	f1-score	support
0	0.99	0.85	0.91	7265
1	0.45	0.92	0.60	971
accuracy			0.86	8236
macro avg	0.72	0.88	0.76	8236
weighted avg	0.92	0.86	0.88	8236

Hình 4.1: Kết quả phân lớp của mô hình Random Forest

	precision	recall	f1-score	support
0	0.92	0.97	0.95	7265
1	0.66	0.39	0.49	971
accuracy			0.90	8236
macro avg	0.79	0.68	0.72	8236
weighted avg	0.89	0.90	0.89	8236

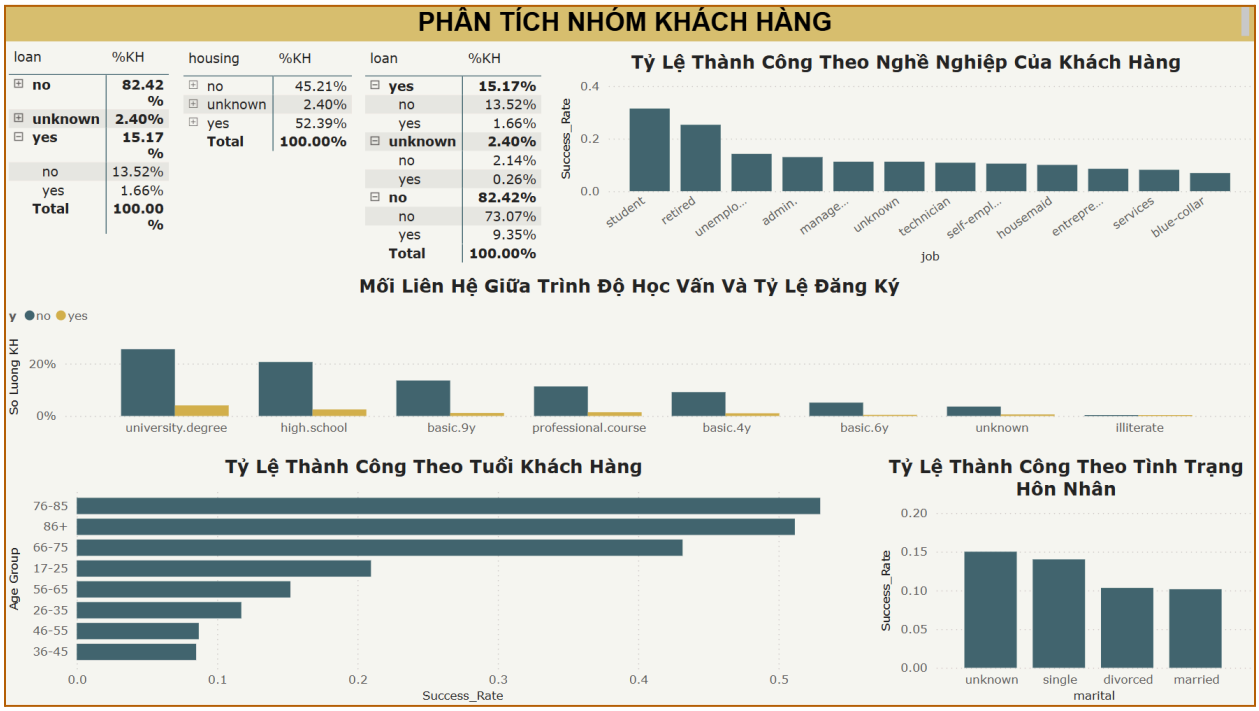
Hình 4.2: Kết quả phân lớp của mô hình Logistic Regression

	precision	recall	f1-score	support
0	0.94	0.95	0.94	7265
1	0.58	0.51	0.54	971
accuracy			0.90	8236
macro avg	0.76	0.73	0.74	8236
weighted avg	0.89	0.90	0.90	8236

Hình 4.3: Kết quả phân lớp của mô hình KNN

Dựa vào kết quả ở hình trên thì mô hình Random Forest có chỉ số Recall ở lớp ‘yes’ cao nhất trong số các mô hình. Vì vậy mô hình Random Forest sẽ được chọn để gán nhãn cho các khách hàng trong tương lai.

5. Dashboard phân tích:



Hình 5.1: Dashboard "Phân tích nhóm khách hàng"

Từ dashboard "Phân tích nhóm khách hàng", ta có thể thấy một số đặc điểm như sau:

➢ **Tỷ lệ thành công theo tuổi khách hàng & theo nghề nghiệp khách hàng:**

- Nhóm khách hàng thuộc thế hệ "baby boomer" và nhóm đối tượng đã nghỉ hưu "retired" (nhóm 66-75 tuổi, nhóm trên 86 tuổi, nhóm 76-85 tuổi) sẽ có tỉ lệ thành công cao nhất.
- Nhóm đối tượng cao thứ hai là nhóm khách hàng thuộc xếp loại "học sinh, sinh viên".

⇒ Có thể tập trung thực hiện các chiến dịch marketing đối với 2 tệp khách hàng này. Hai tệp này có tỉ lệ thành công cao do, tệp khách hàng lớn tuổi/đã nghỉ hưu sẽ có xu hướng thế chấp nhà cửa để lãnh khoản lương hưu; còn đối với tệp khách trẻ hơn họ sẽ có xu hướng vay nợ sinh viên để có thể chi trả cho đại học, cao đẳng.

➢ **Tỷ lệ thành công theo tình trạng hôn nhân:**

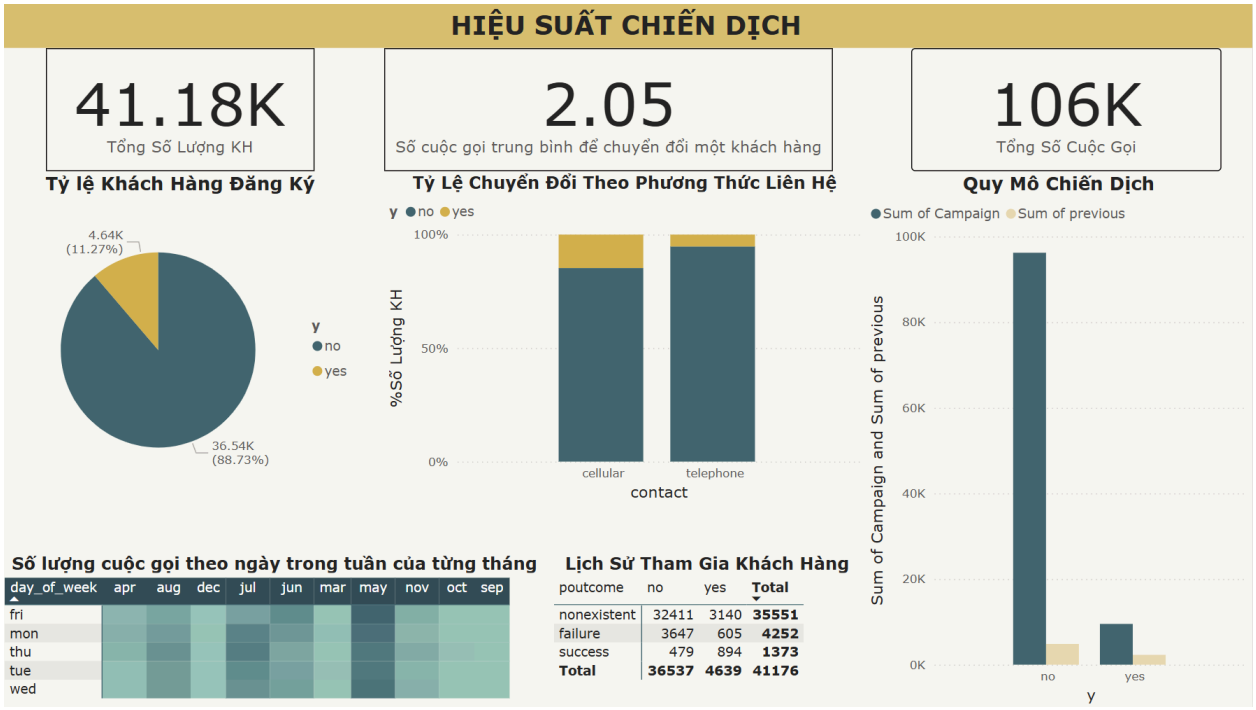
- Đối với nhóm khách hàng còn độc thân, tỷ lệ thành công đối với tệp này khá ca.
- Ngược lại, 2 nhóm "đã ly dị" và "đã kết hôn" sẽ có tỷ lệ thấp hơn đáng kể.

➢ **Trình độ học vấn cao có mối tương quan với tỷ lệ đăng ký cao**

- Những người có bằng đại học (university degree) và trung học phổ thông (high school) chiếm tỷ lệ khách hàng lớn hơn so với nhóm có trình độ học vấn thấp.
- Điều này có thể cho thấy những người có trình độ cao hơn dễ tiếp nhận thông tin và có nhu cầu cao hơn.

⇒ Tạo nội dung truyền thông phù hợp với từng nhóm học vấn để tiếp cận hiệu quả hơn.

Đặc điểm khách hàng	**Chi tiết khách hàng**
Độ tuổi	• Nhóm baby boomer (66^{+} tuổi) • Nhóm tuổi thiếu niên (17-25 tuổi)
Nghề nghiệp	Hai nhóm đối tượng: học sinh sinh viên (student) và nhóm đã nghỉ hưu (retired).
Học vấn	Nhóm đối tượng có học vấn cao, trường đại học và THPT (university degree và high school).
Tình trạng hôn nhân	Tập trung vào tệp đối tượng còn độc thân.



Dựa vào biểu đồ chúng ta có những góc nhìn như sau:

- Tỷ lệ Khách Hàng Đăng Ký: 41.18K, tương đương 11.27% tổng số lượng khách hàng tiềm năng.

- Số Cuộc Gọi Trung Bình để Chuyển Đổi: 2.05 cuộc gọi để chuyển đổi một khách hàng.

- Tổng Số Cuộc Gọi: 106K cuộc gọi.