MÔN KHAI PHÁ DỮ LIỆU

PHÂN TÍCH HÀNH VI KHÁCH HÀNG VÀ TỐI ƯU HÓA CHIẾN DỊCH TELEMARKETING

cho sản phẩm tiền gửi có kỳ hạn tại ngân hàng

Nhóm: 10
GVHD: TS. Nguyễn Thành Huy

TỔNG QUAN ĐỒ ÁN

Thực trạng:

- Khối lượng cuộc gọi khổng lồ (>7.000 cuộc/tháng)
- Tỷ lệ thành công thấp

Mục tiêu nghiên cứu:

- Làm thế nào để tối ưu hóa danh sách gọi?
- Làm sao để đạt hiệu quả cao nhất với nguồn lực thấp nhất?
- Tối đa hóa số lượng khách hàng đăng ký tiền gửi có kỳ hạn ?

Bài toán:

- Logistic Regression/Decision Tree/Random Forest/Gradient Boosting
- K-Means Clustering

Trực quan hóa:

- Dashboard tương tác bằng Streamlit

TỔNG QUAN ĐỒ ÁN

Conversion Rate là tỷ lệ khách hàng thực hiện hành động mong muốn trên tổng số khách hàng được tiếp cận.

$$\text{Conversion Rate} = \frac{\text{Số khách hàng chuyển đổi}}{\text{Tổng số khách hàng được tiếp cận}}$$

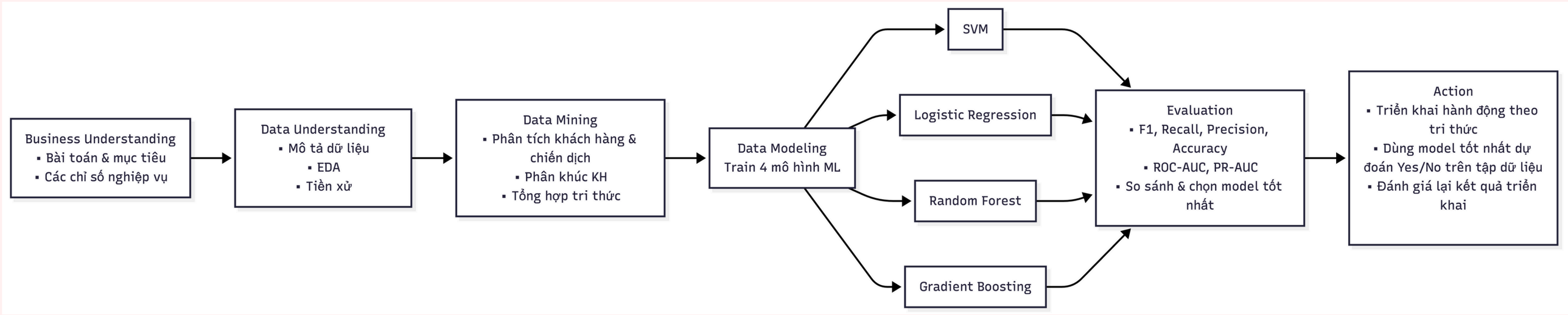
Precision@X% → Trong số khách hàng được gọi, bao nhiêu % sẽ nói **YES**

$$\boxed{\text{Lift@TopX\%} = \frac{\text{Conversion Rate@TopX\%}}{\text{Conversion Rate (toàn bộ tập)}}}$$

Lift@X% → Hiệu quả tốt hơn bao nhiêu lần so với gọi ngẫu nhiên

$$\boxed{\text{Lift@TopX\%} = \frac{\text{Conversion Rate@TopX\%}}{\text{Conversion Rate (toàn bộ tập)}}}$$

Quy trình thực hiện



MÔ TẢ DỮ LIỆU

Nguồn & Kích thước:

- Dữ liệu: Bank Marketing Dataset (45.211 quan sát, 17 biến)
- Loại biến: 7 biến số và 10 biến phân loại

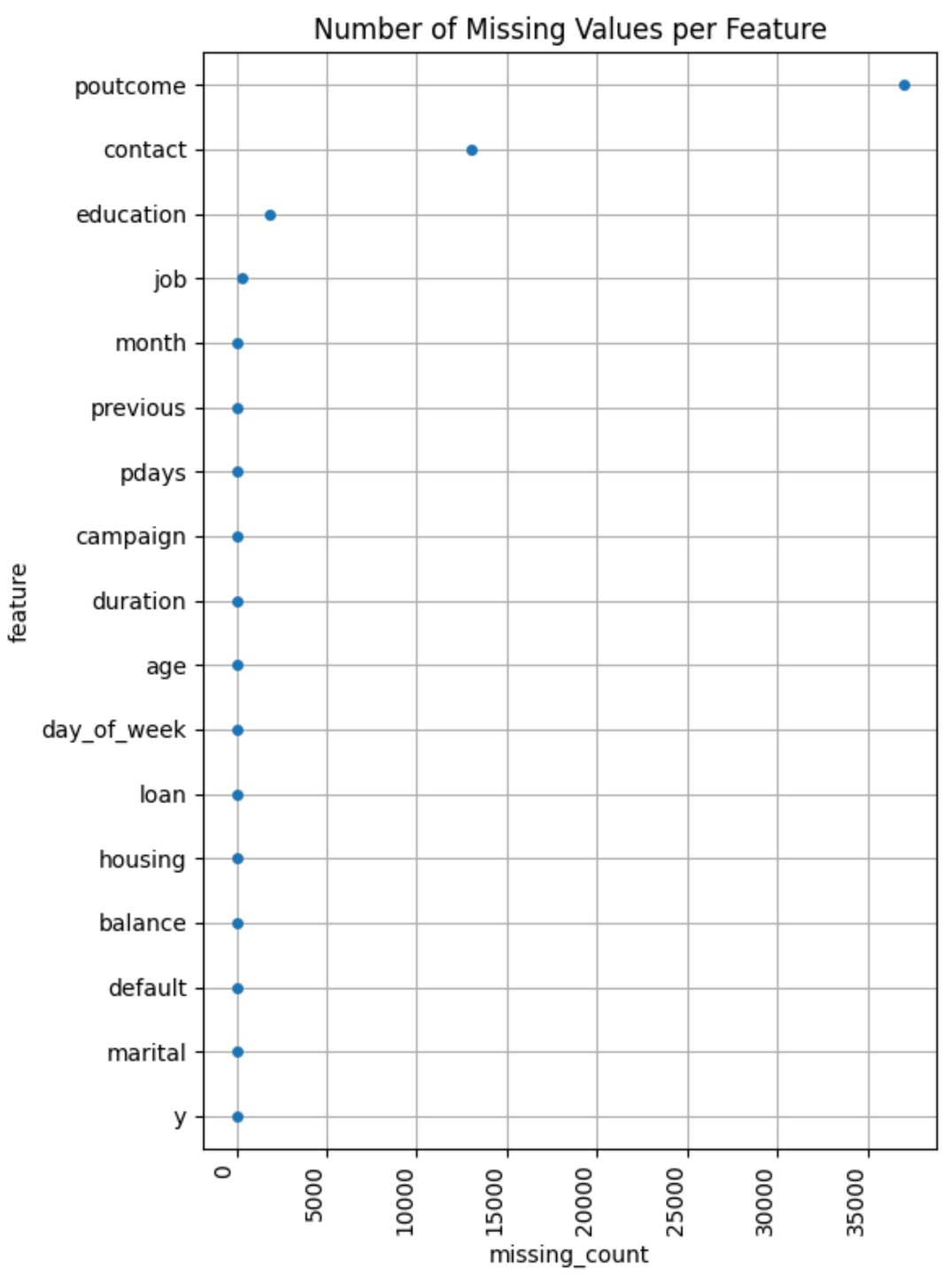
Các nhóm biến chính:

- Thông tin cá nhân: *age*, *job*, *marital*, *education*
- Tài chính: *balance*, *housing*, *loan*, *default*
- Chiến dịch: *contact*, *day*, *month*, *campaign*, *poutcome*

Data Quality

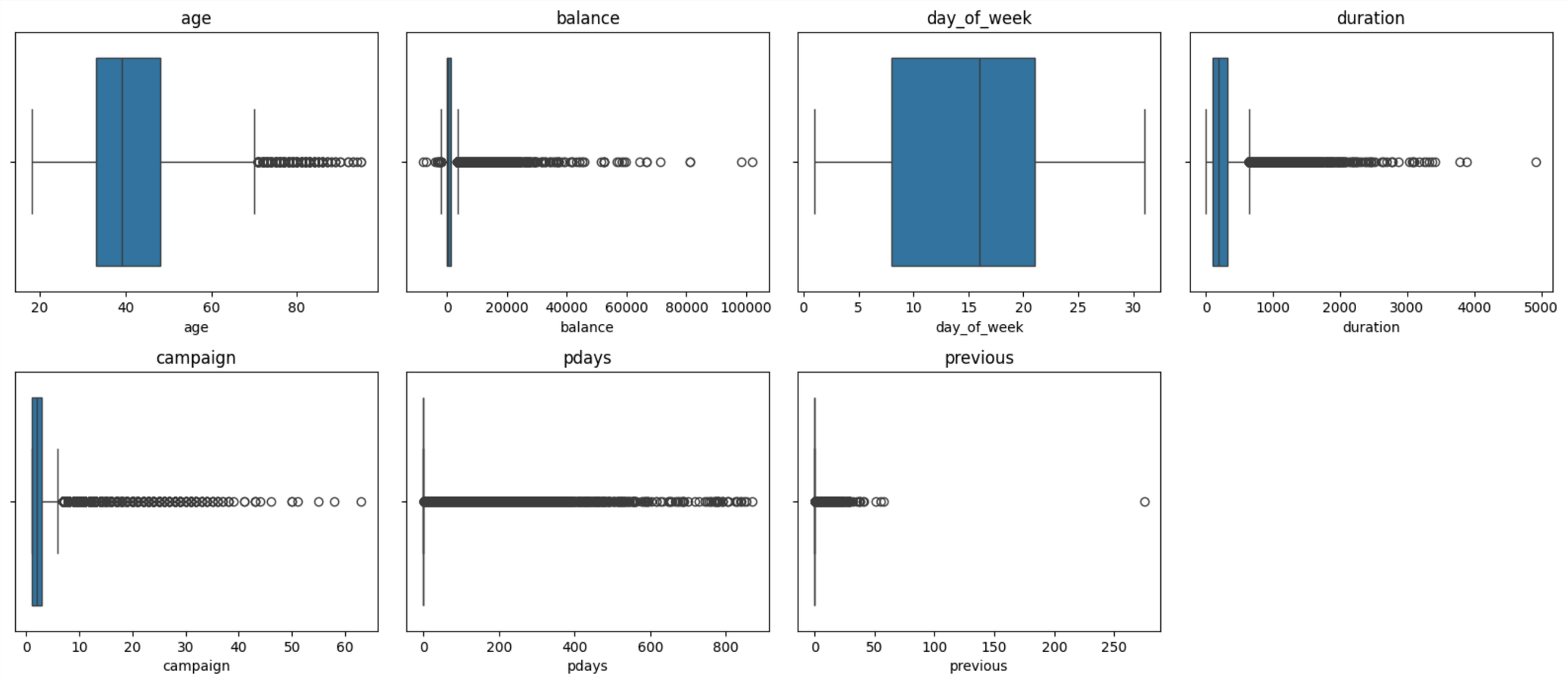
Missing values - Giữ lại và xử lý như một nhóm riêng biệt (“Unknown")

Xử lý giá trị trùng lặp



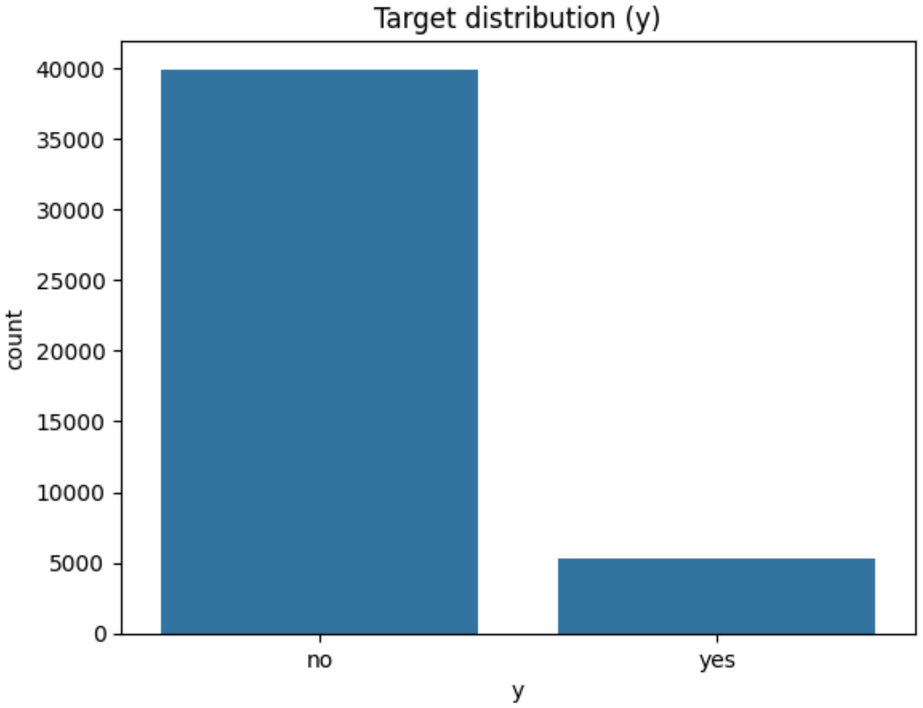
Data Quality

Giá trị ngoại lai (Outliers)- Phản ánh khách hàng VIP (số dư lớn) hoặc khách hàng khó tính (gọi nhiều lần) → Giữ nguyên

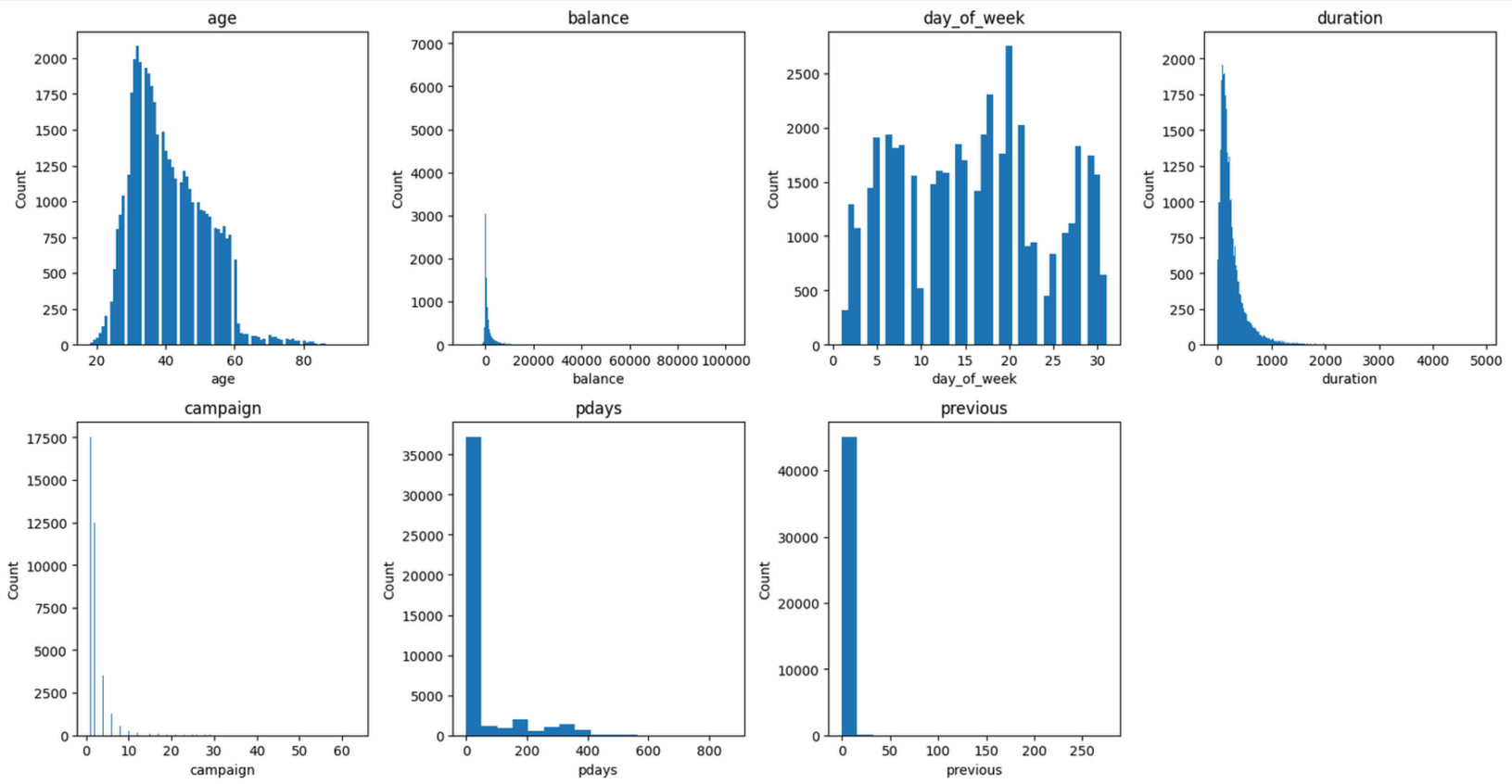


EDA

MẤT CÂN BẰNG DỮ LIỆU



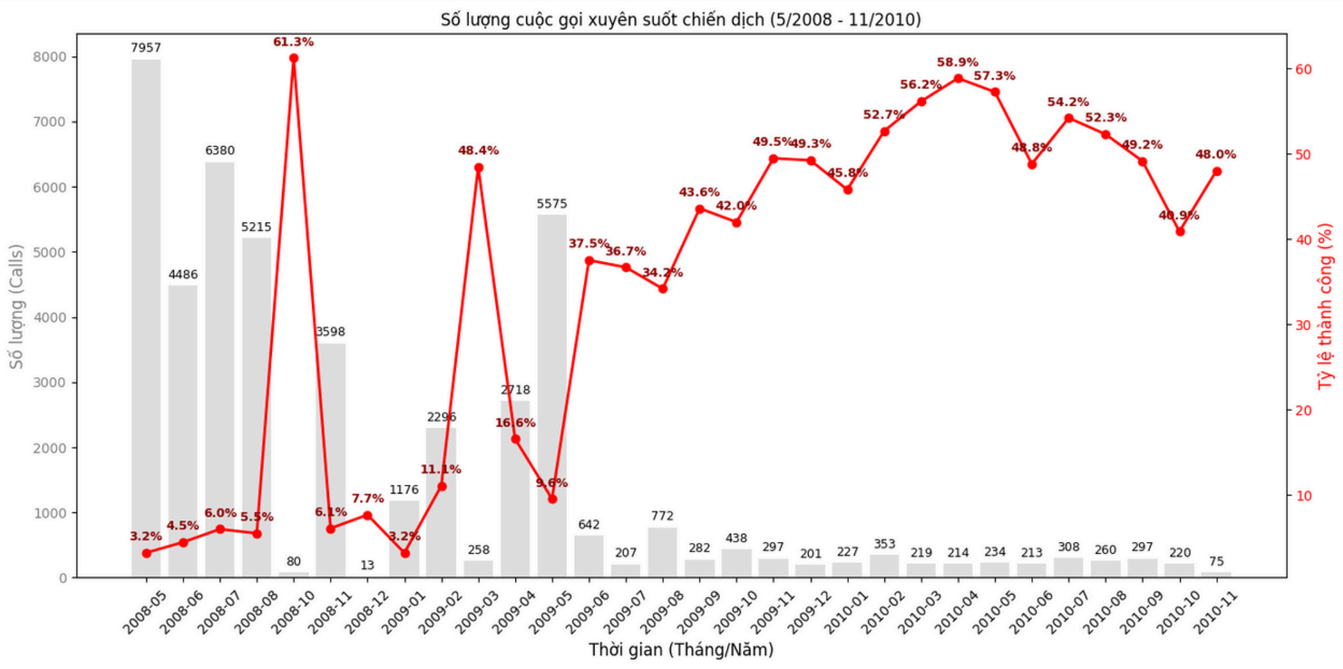
Xử lý bằng SMOTE và **Class-weight**



PHÂN PHỐI CÁC BIẾN SỐ VẪN CÒN LỆCH (LOG-TRANSFORM)

Chiến dịch & Vận hành (Campaign Analysis)

CÁC YẾU TỐ LỊCH SỬ (KẾT QUẢ CŨ) VÀ VẬN HÀNH (THỜI LƯỢNG, TẦN SUẤT GỌI) ẢNH HƯỞNG THẾ NÀO ĐẾN TỶ LỆ CHỐT ĐƠN?



3 giai đoạn của chiến dịch

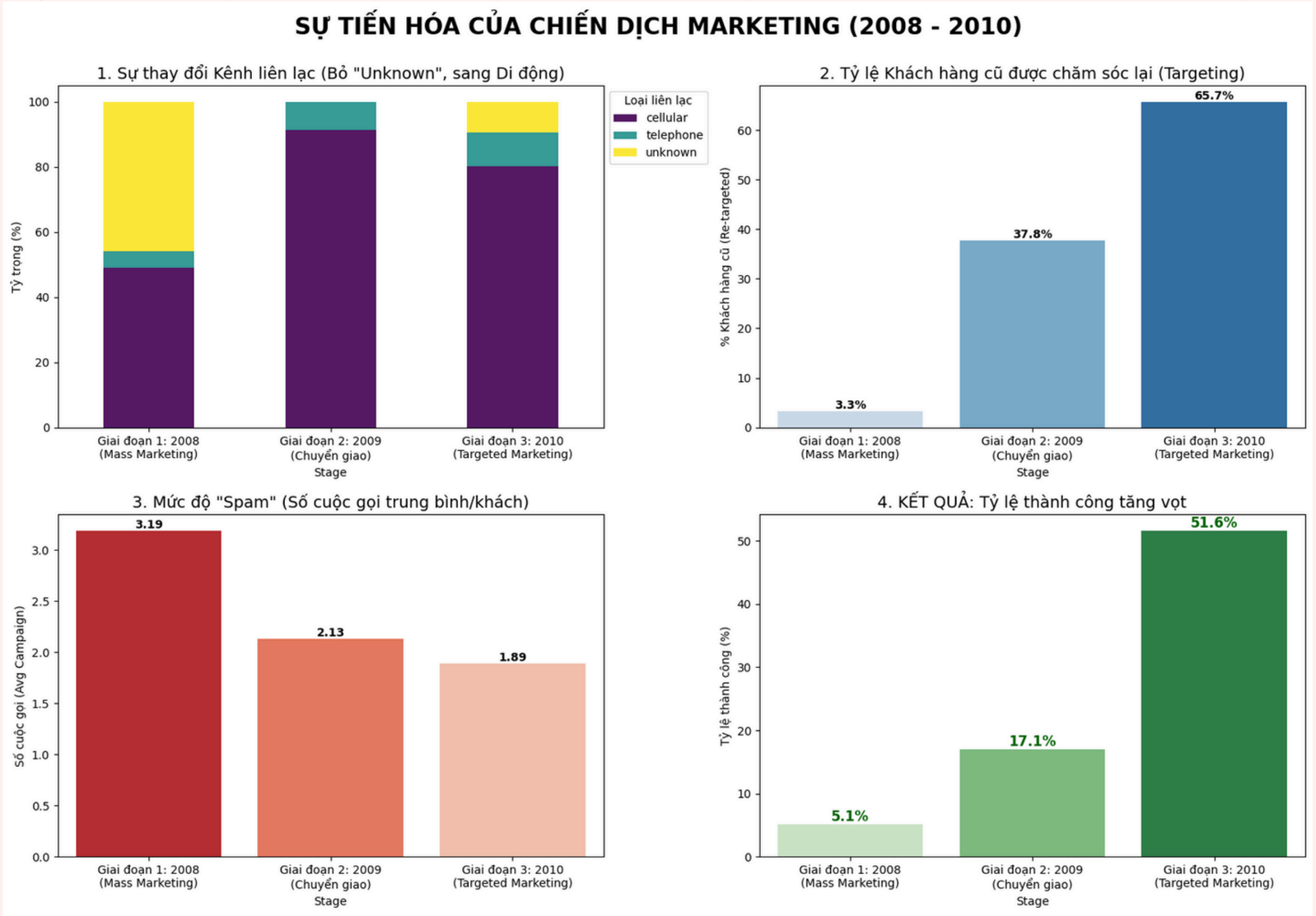
Giai đoạn 1: Mass Marketing

Giai đoạn 2: Chuyển giao

Giai đoạn 3: Targeted Marketing

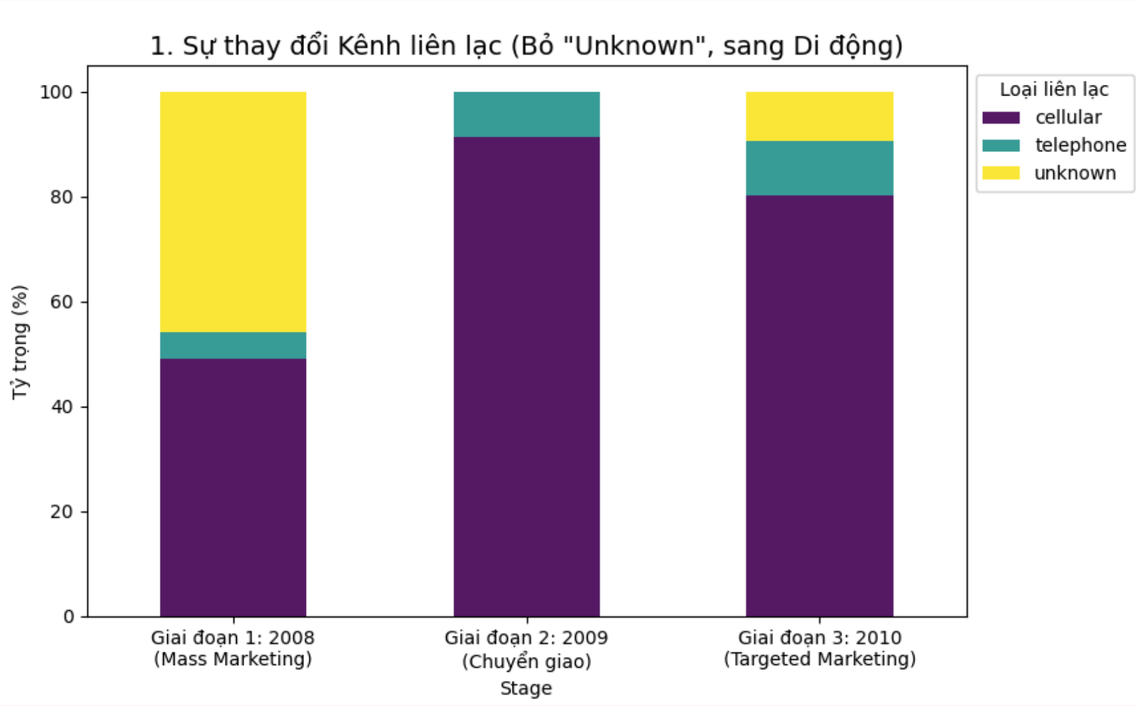
Chiến dịch & Vận hành (Campaign Analysis)

CÁC YẾU TỐ LỊCH SỬ (KẾT QUẢ CŨ) VÀ VẬN HÀNH (THỜI LƯỢNG, TẦN SUẤT GỌI) ẢNH HƯỞNG THẾ NÀO ĐẾN TỶ LỆ CHỐT ĐƠN?



Chiến dịch & Vận hành (Campaign Analysis)

CÁC YẾU TỐ LỊCH SỬ (KẾT QUẢ CŨ) VÀ VẬN HÀNH (THỜI LƯỢNG, TẦN SUẤT GỌI) ẢNH HƯỞNG THẾ NÀO ĐẾN TỶ LỆ CHỐT ĐƠN?



Tối ưu hóa danh sách khách hàng

Loại bỏ khách hàng có kênh liên lạc không xác định/thiếu số điện thoại di động khỏi danh sách tác nghiệp

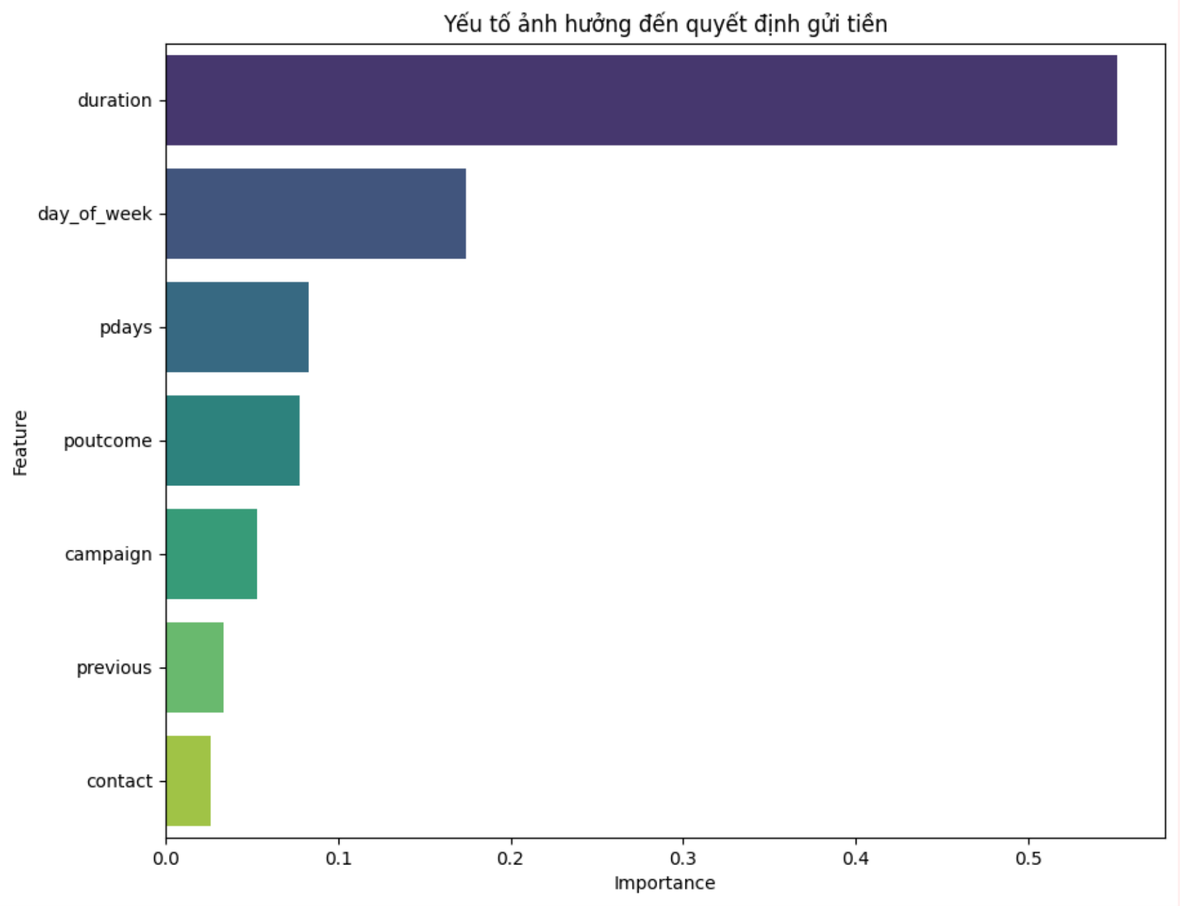
Chiến dịch & Vận hành

CÁC YẾU TỐ LỊCH SỬ (KẾT QUẢ CŨ) VÀ VẬN HÀNH (THỜI LƯỢNG, TẦN SUẤT GỌI) ẢNH HƯỞNG THẾ NÀO ĐẾN TỶ LỆ CHỐT ĐƠN?

Random Forest - Feature Importance

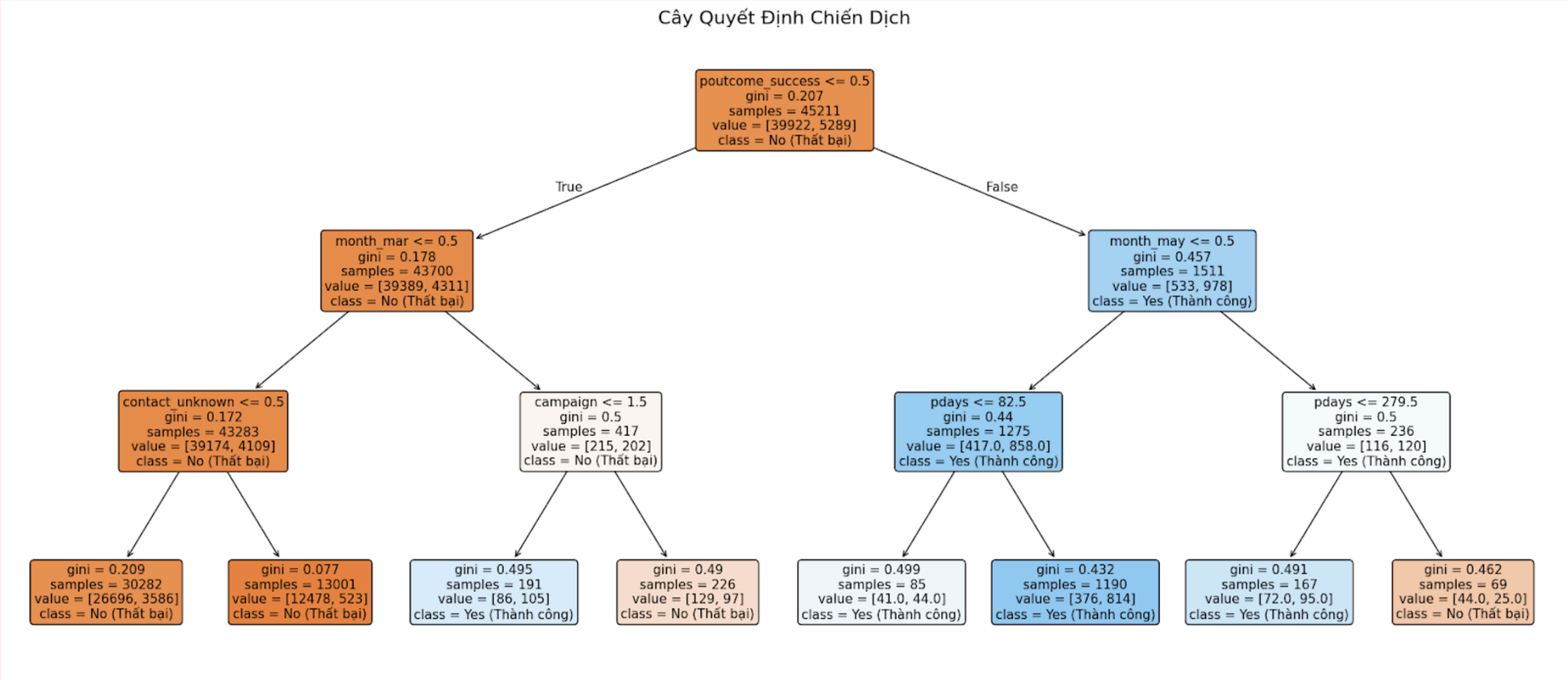
- Biến duration có kết quả importance cao nhất
- Không thể chủ động kiểm soát thời lượng cuộc gọi
- Duration dài → Đồng ý → Data leakage

→ Dùng duration để dự báo kết quả là vô nghĩa



Chiến dịch & Vận hành

CÁC YẾU TỐ LỊCH SỬ (KẾT QUẢ CŨ) VÀ VẬN HÀNH (THỜI LƯỢNG, TẦN SUẤT GỌI) ẢNH HƯỞNG THẾ NÀO ĐẾN TỶ LỆ CHỐT ĐƠN?

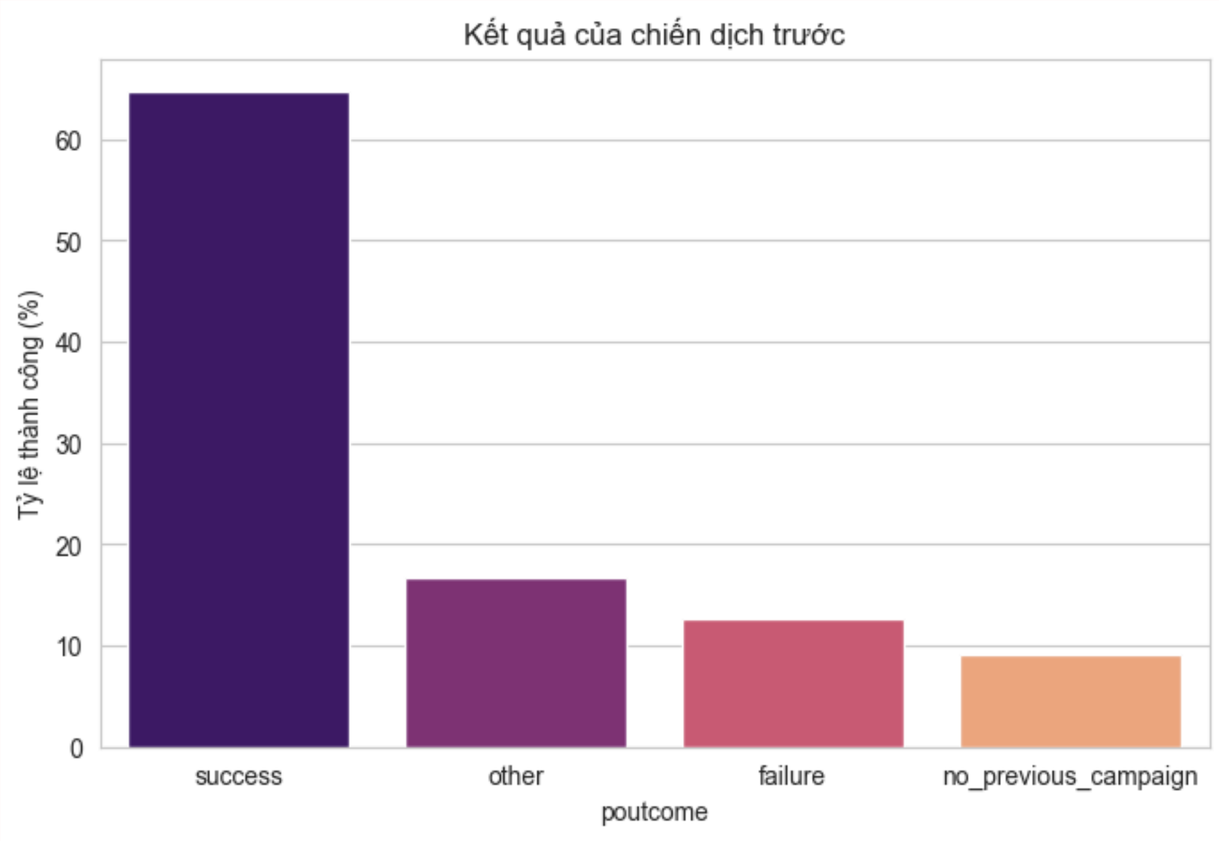


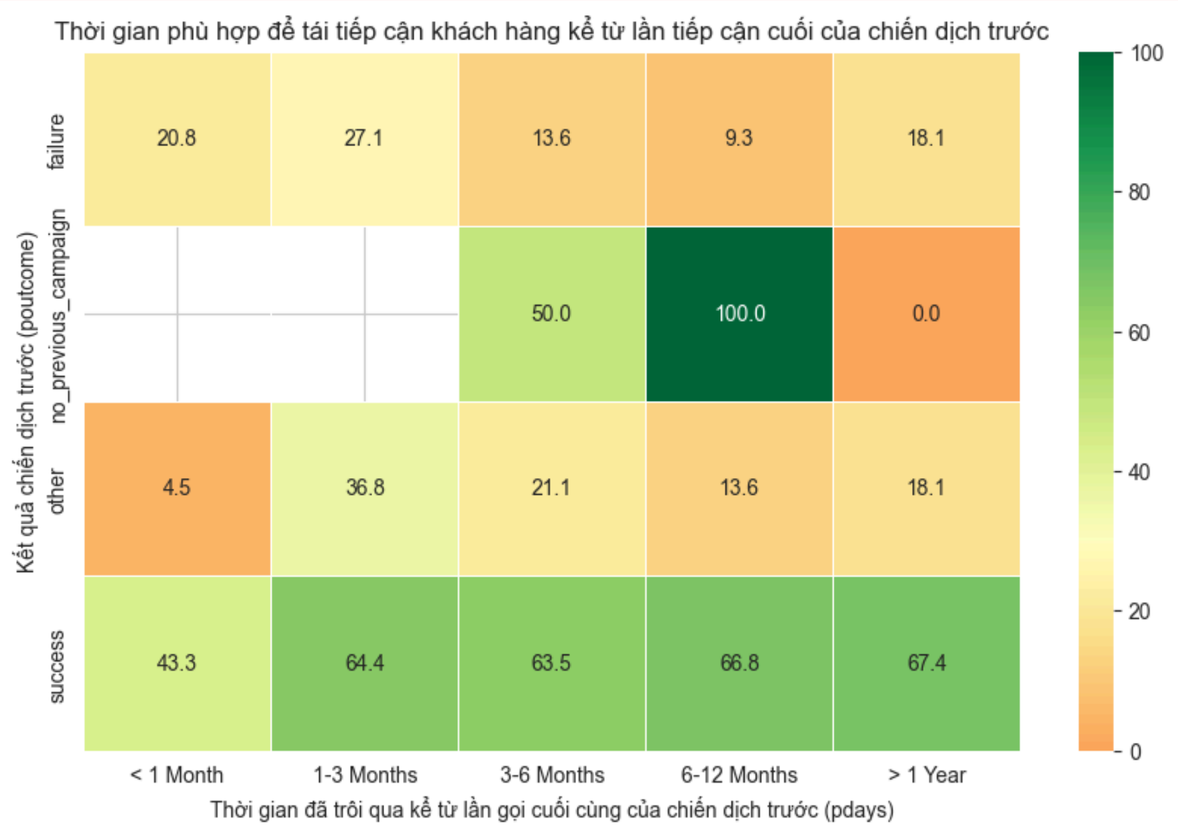
Decision Tree

Cây quyết định cho thấy poutcome
là yếu tố phân tách quan trọng nhất

Chiến dịch & Vận hành

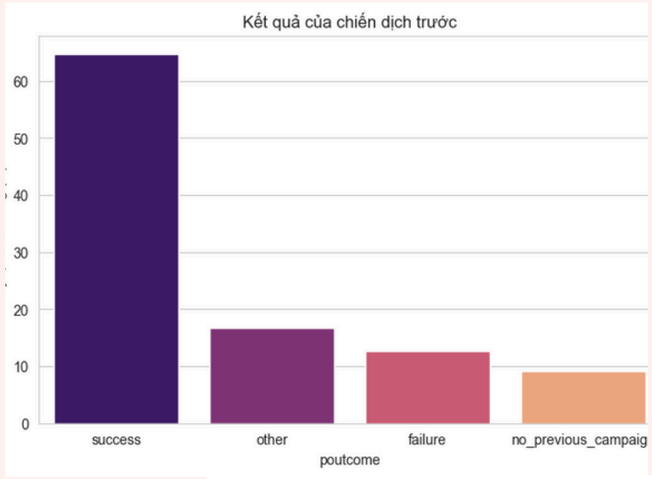
CÁC YẾU TỐ LỊCH SỬ (KẾT QUẢ CŨ) VÀ VẬN HÀNH (THỜI LƯỢNG, TẦN SUẤT GỌI) ẢNH HƯỞNG THẾ NÀO ĐẾN TỶ LỆ CHỐT ĐƠN?

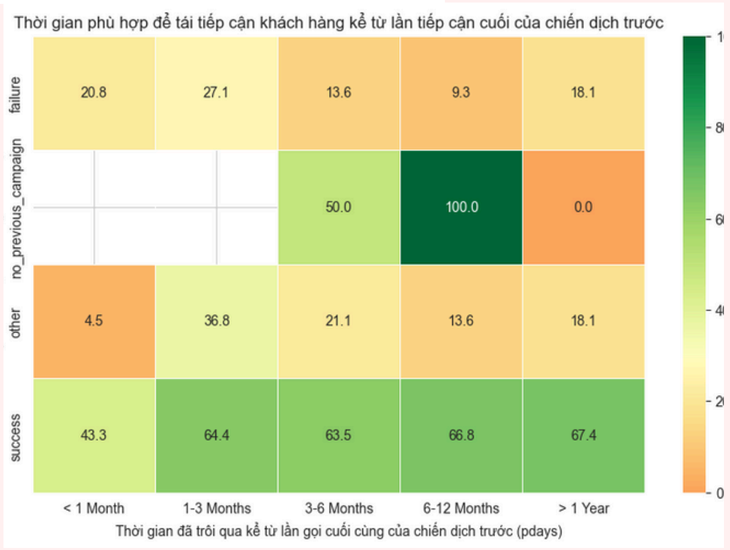




Chiến dịch & Vận hành

CÁC YẾU TỐ LỊCH SỬ (KẾT QUẢ CŨ) VÀ VẬN HÀNH (THỜI LƯỢNG, TẦN SUẤT GỌI) ẢNH HƯỞNG THẾ NÀO ĐẾN TỶ LỆ CHỐT ĐƠN?



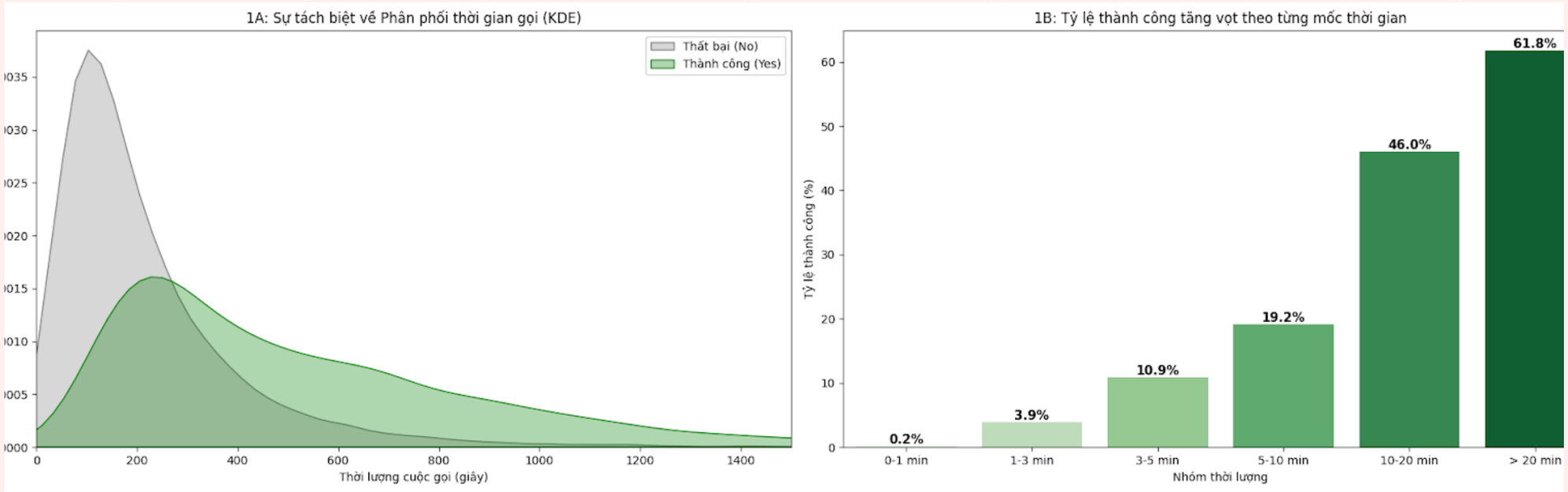


Quy tắc vận hành:

- Gán nhãn VIP và **gọi ngay lập tức** đối với những khách hàng đã có **lịch sử success**
- Liên lạc KH có thời gian chờ nằm trong khoảng từ **6 đến 12 tháng** kể từ lần liên hệ trước
- Liên lạc lại những người có kết quả failure lần trước có thời gian chờ từ 1 đến 3 tháng

Chiến dịch & Vận hành

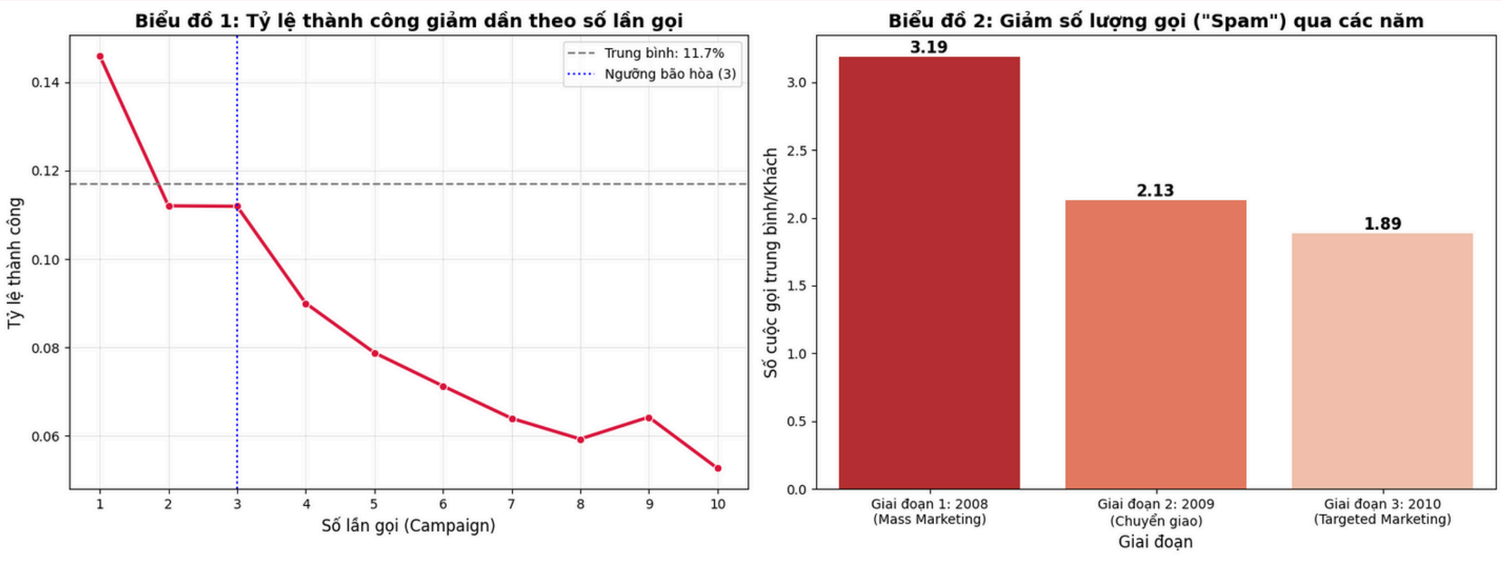
CÁC YẾU TỐ LỊCH SỬ (KẾT QUẢ CŨ) VÀ VẬN HÀNH (THỜI LƯỢNG, TẦN SUẤT GỌI) ẢNH HƯỞNG THẾ NÀO ĐẾN TỶ LỆ CHỐT ĐƠN?



- Các cuộc gọi từ chối kết thúc trong khoảng 60–100 giây.
- Các cuộc gọi thành công có giá trị trung vị thường vượt 200 giây.
- **Mốc 120 giây** là một ngưỡng chuyển tiếp quan trọng
- Khách hàng đã chuyển sang giai đoạn cân nhắc

Chiến dịch & Vận hành

CÁC YẾU TỐ LỊCH SỬ (KẾT QUẢ CŨ) VÀ VẬN HÀNH (THỜI LƯỢNG, TẦN SUẤT GỌI) ẢNH HƯỞNG THẾ NÀO ĐẾN TỶ LỆ CHỐT ĐƠN?



Quy tắc vận hành:

- Liên lạc hơn 2 lần → Cân nhắc ngưng chiến dịch với khách đó

- Tỷ lệ chuyển đổi đạt đỉnh ở **cuộc gọi thứ nhất và thứ hai**

- Sau đó giảm sâu tại cuộc gọi thứ ba và gần như tiệm cận mức 0 từ lần liên hệ thứ tư trở đi

Khách hàng

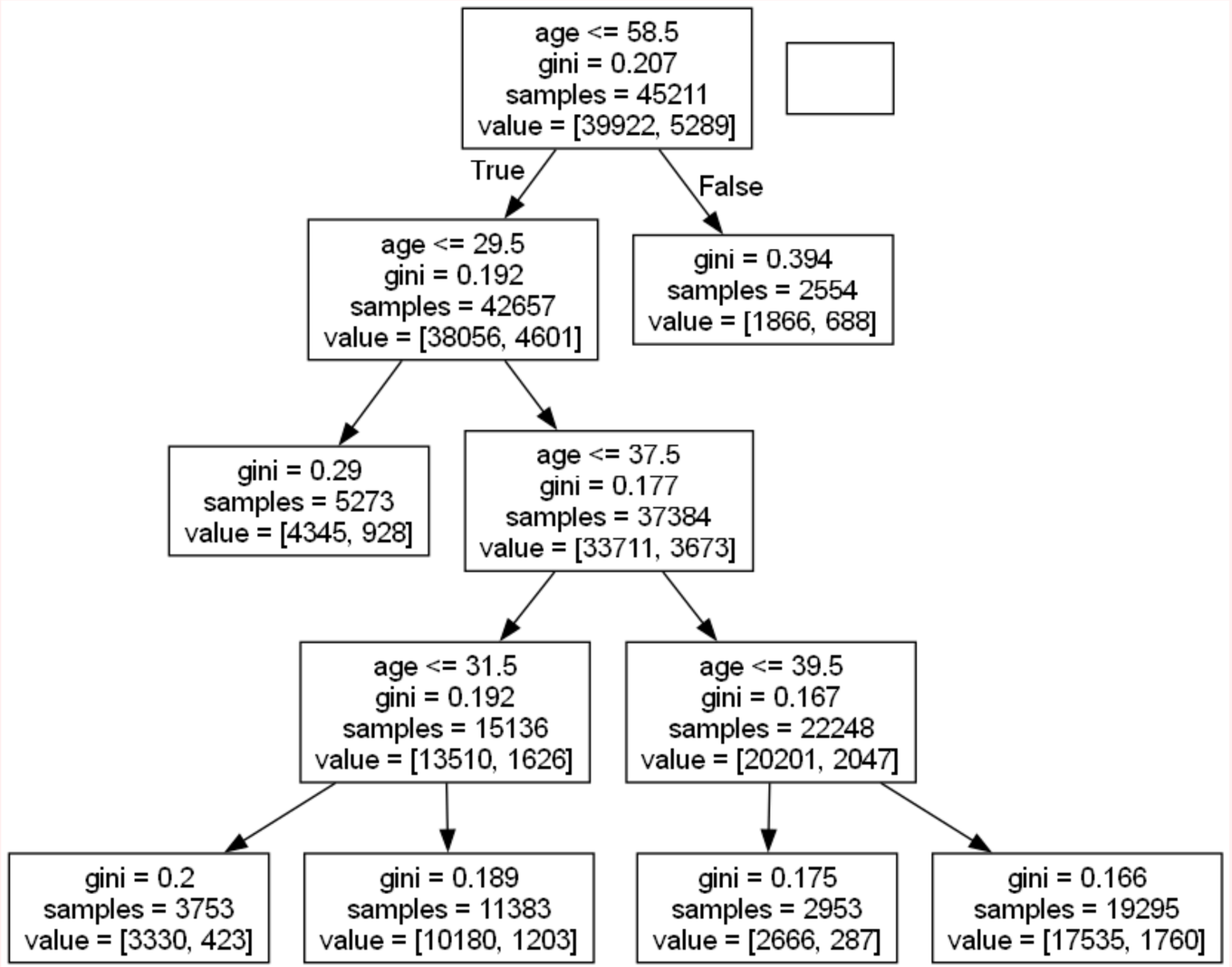
Các nhóm tuổi (Decision Tree)

Nhóm 1: từ 0–29 tuổi (12%)

Nhóm 2: từ 29–37 tuổi (33%)

Nhóm 3: từ 37–58 tuổi (49%)

Nhóm 4: trên 58 tuổi (6%)



Khách hàng

Các nhóm số dư tài khoản (Optimal Binning)

Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
(-inf, -46.50)	3193	0.070624	3027	166	0.051989	0.882041	0.039195	0.004747
[-46.50, 60.50)	7628	0.16872	7034	594	0.077871	0.450333	0.02877	0.003566
[60.50, 798.50)	17577	0.388777	15614	1963	0.11168	0.052396	0.001046	0.000131
[798.50, 1578.50)	6369	0.140873	5521	848	0.133145	-0.147865	0.003259	0.000407
[1578.50, inf)	10444	0.231006	8726	1718	0.164496	-0.396152	0.042091	0.005227

Khách hàng

Tầm quan trọng của biến

Biến	Importance
age_group	0.38
housing	0.378
balance_group	0.158
loan	0.081
job	0.003
marital	0.002
education	0
default	0

Tương tác biến

	job	age_group	balance_group	housing	loan
job	-	0.119	0.093	0.084	0.077
age_group	0.119	-	0.103	0.05	0.055
balance_group	0.094	0.103	-	0.067	0.012
housing	0.084	0.05	0.067	-	0.031
loan	0.077	0.056	0.012	0.032	-

Khách hàng

Phương pháp:

- Hàm sinh luật
- Quan sát biểu đồ
- Xác nhận lại bằng Decision Tree

- *Số lượng mẫu > 300 (support)*
- *Thuộc 25% tỷ lệ đăng ký cao nhất (quantile)*

Kết luận:

- Việc vay mua nhà là rào cản đối với hành vi tiết kiệm
- Khách hàng trẻ (<29) và lớn tuổi (>58) có xác suất đăng ký cao hơn nhóm trung niên (29–58), bất kể nghề nghiệp
- Nghề nghiệp không có tác động độc lập mà phụ thuộc vào bối cảnh tài chính

Xác định nhóm khách hàng tiềm năng

tăng Conversion Rate và ROI bằng cách ưu tiên nhóm khách có xác suất đăng ký cao.

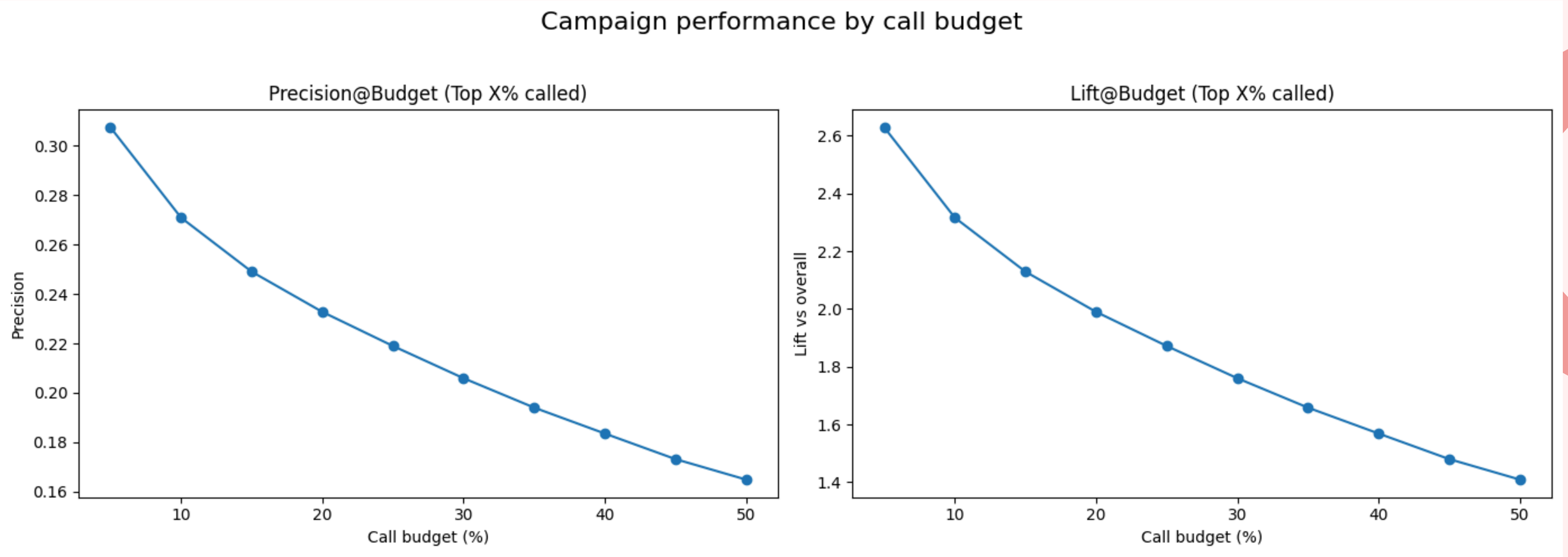
Phương pháp:

- So sánh mô hình: Logistic Regression vs Decision Tree vs Dummy
- Chọn Logistic Regression (hiệu năng tổng quát tốt, ổn định)
- Đánh giá theo xếp hạng: Precision@K & Lift@K (phù hợp call budget)
- Chọn ngưỡng (threshold) theo F1 để cân bằng bắt đúng và giảm gọi nhầm (FP)

Bảng chiến lược tiếp cận theo thứ hạng khách hàng

Phân lớp (Rank)	Tỷ lệ phản hồi (Precision)	Hiệu quả so với ngẫu nhiên (Lift)	Hành động đề xuất
Top 5%	29.40%	**2.52 lần**	Gọi điện trực tiếp ngay lập tức
Top 10-20%	~24%	**~2.1 lần**	Gửi SMS/Email kèm ưu đãi
Top 30%	<20%	<1.7 lần	Lưu hồ sơ, chưa cần tiếp cận

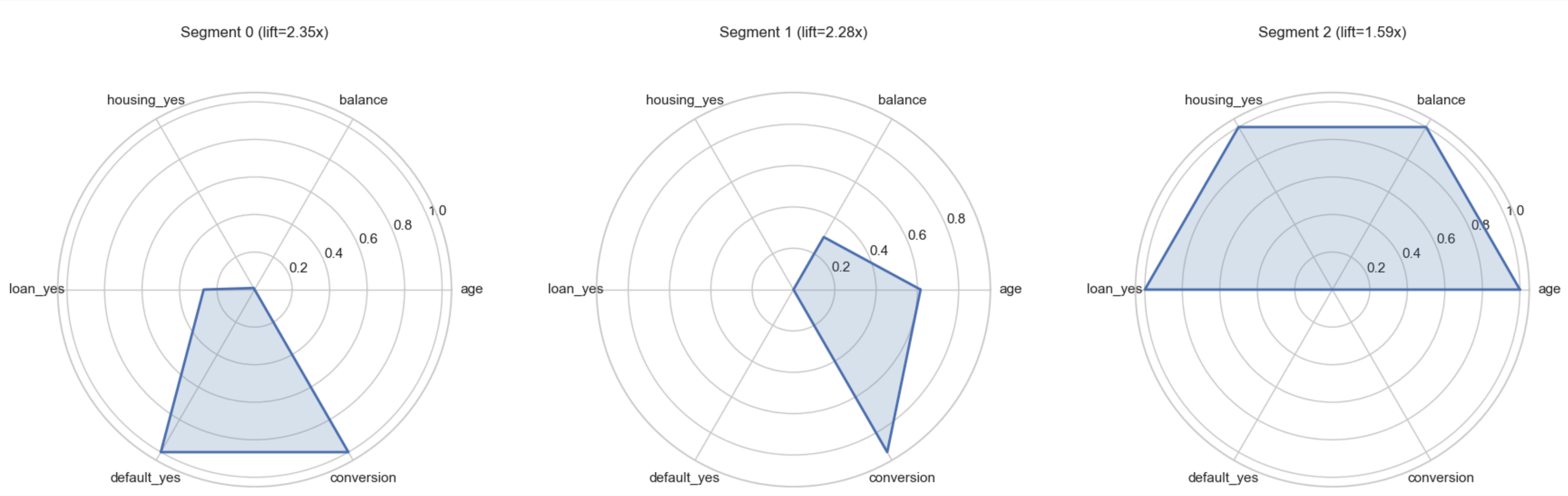
Hiệu quả Chiến dịch theo Ngưỡng Ngân sách



ngưỡng ngân sách khoảng 10%–20% là vùng hợp lý, trong đó 10% là lựa chọn ưu tiên khi nguồn lực hạn chế.

Chân dung phân khúc khách hàng tiềm năng (Top 10%)

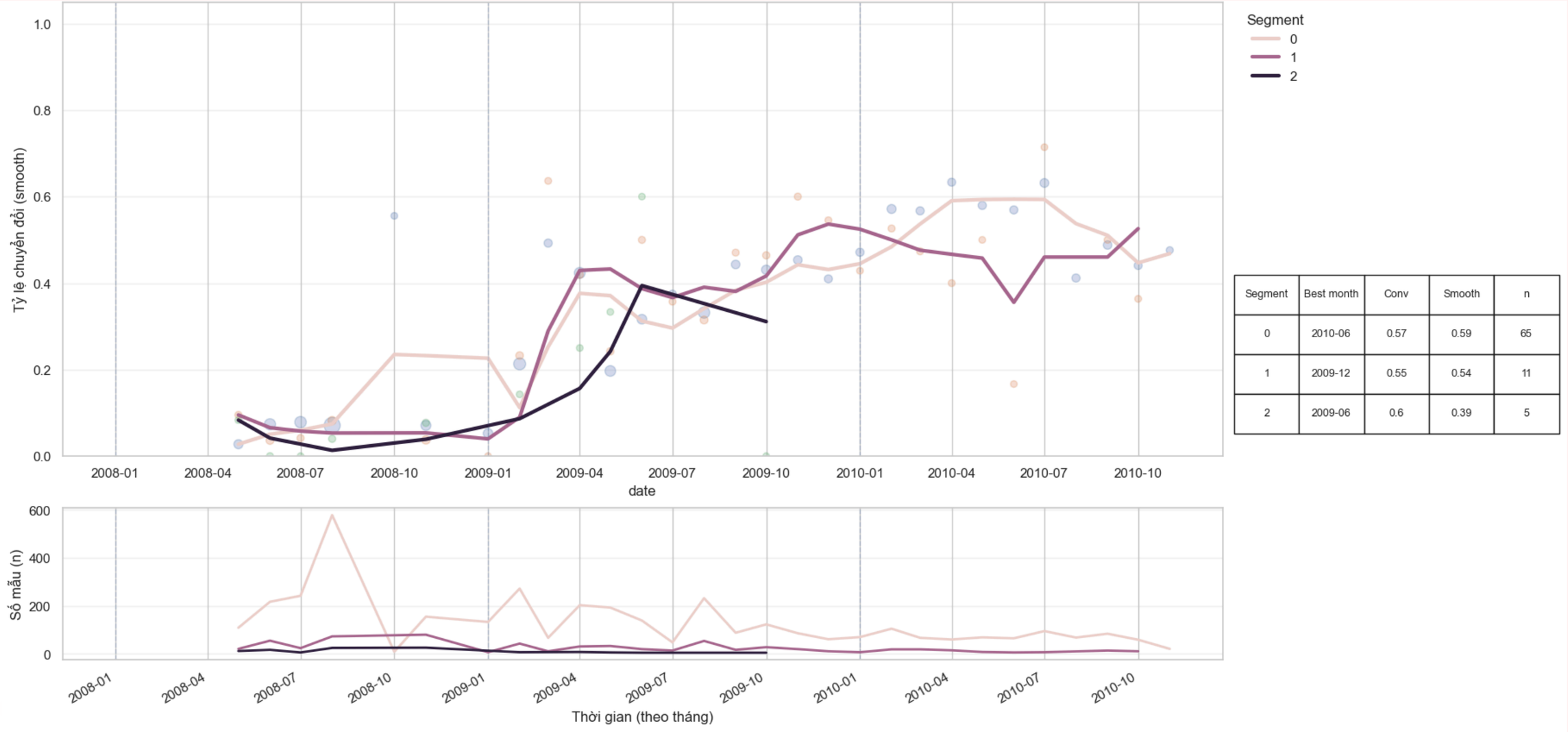
Phân khúc khách hàng ngách với MiniBatchKMeans (K=3) để phân nhóm khách hàng tiềm năng với các features (pre-call): age, balance, (previous/y_yes), housing, pdays_contacted.



Radar thể hiện chân dung mô tả theo phân khúc (profile summary).
conversion dùng để đánh giá hiệu quả theo phân khúc, không đưa vào clustering (tránh leakage).

Chiến lược gọi theo phân khúc khách hàng -when to call?

Xu hướng conversion theo thời gian cho từng phân khúc và số lượng quan sát theo tháng



Segment	Best month	Conv	Smooth	n
0	2010-06	0.57	0.59	65
1	2009-12	0.55	0.54	11
2	2009-06	0.6	0.39	5

Chiến lược gọi theo phân khúc khách hàng

Segment 0 – Mass call (chủ lực)

Chân dung: Trẻ, độc thân; ít nợ xấu/ít vay; số dư thấp–TB

Thời điểm: Ưu tiên tháng 4–6

Hành động: Gọi diện rộng, tối ưu sản lượng

Segment 1 – Selective call (hiệu quả/call)

Chân dung: Trung niên; nhiều quản lý; số dư cao; ít vay

Thời điểm: Ưu tiên tháng 10–12

Hành động: Gọi chọn lọc + tư vấn mục tiêu tài chính

Segment 2 – VIP / 1-1 (không đại trà)

Chân dung: Lớn tuổi; số dư rất cao; đa số đã kết hôn; tỷ lệ vay cao hơn

Thời điểm: Không theo mùa cố định

Hành động: Chỉ tiếp cận khi có gói/ưu đãi phù hợp (VIP/wealth)

Lựa chọn khách hàng

- Khách hàng chưa có khoản vay mua nhà, dưới 30 tuổi độc thân, gọi các tháng 4, 5 và 6
- Khách hàng trên 60 tuổi, đã kết hôn, tài chính ổn định cũng, chuẩn bị hình thức tư vấn cá nhân hóa
- Khách hàng trung niên, làm việc trong lĩnh vực quản lý, tài chính tốt, gọi có chọn lọc vào các tháng cuối năm
- Ưu tiên các khách hàng có số điện thoại di động cá nhân, có lần liên hệ gần nhất từ 90–180 ngày

Tối ưu hóa ngân sách: có thể loại bỏ các khách hàng có số dư tài khoản dưới mức trung bình

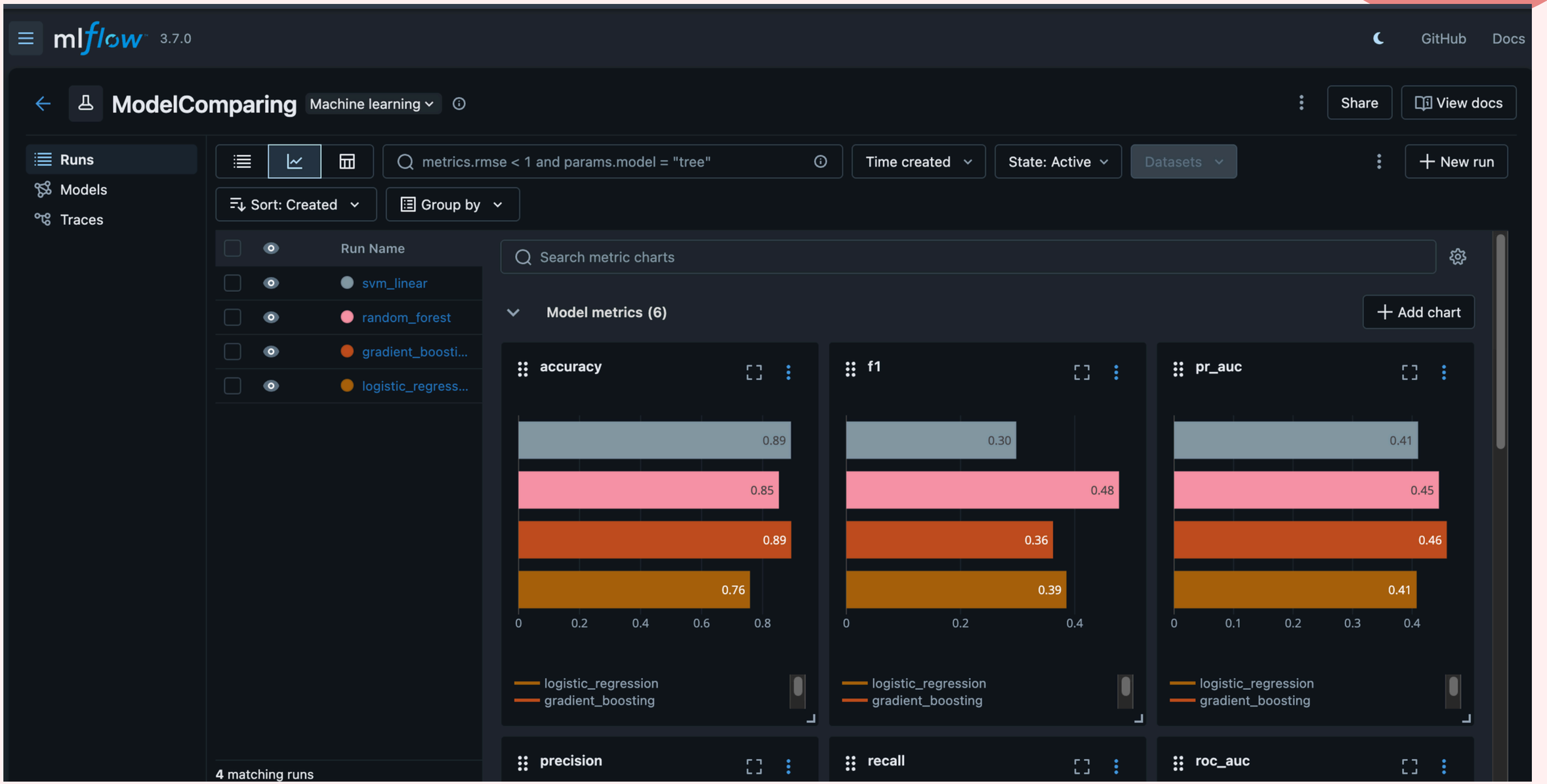
Triển khai chiến dịch

- Liên lạc khách hàng hơn hai lần → cân nhắc loại khỏi danh sách liên hệ
- Thời lượng cuộc gọi đạt từ 120 giây trở lên → nhân viên đẩy mạnh tư vấn và khai thác nhu cầu để tăng khả năng chốt giao dịch

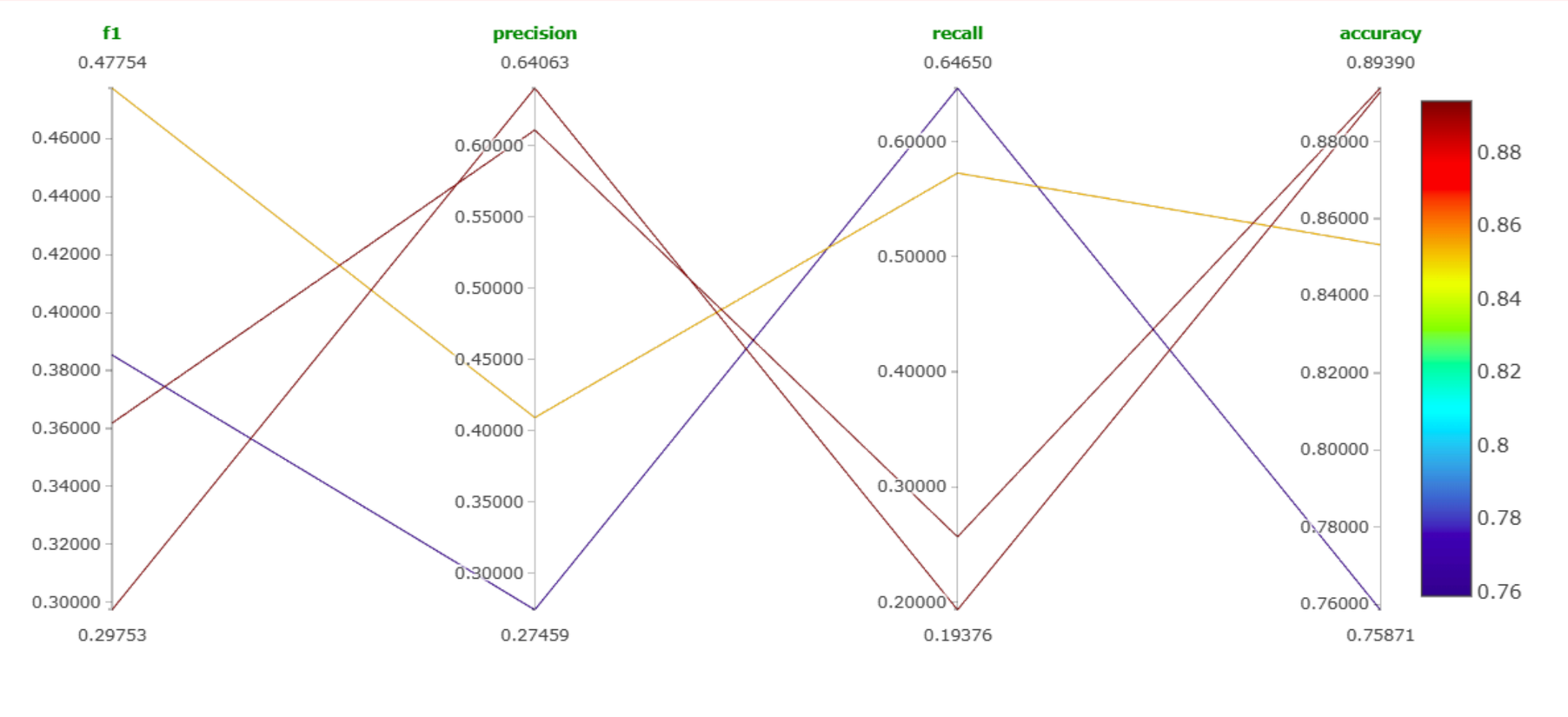
MÔ HÌNH VÀ PHƯƠNG PHÁP

	SVM	Random Forest	Logistic Regression	Gradient Boosting
Lý do chọn	Giúp giảm gọi nhầm khách hàng không tiềm năng nhờ precision cao.	Bắt tốt quan hệ phi tuyến và tương tác biến, hỗ trợ phân tích hành vi khách hàng.	Phù hợp xếp hạng khách hàng theo xác suất, dễ tối ưu chi phí gọi.	Phù hợp với cách tiếp cận xếp hạng và ưu tiên khách hàng để tối ưu chi phí (xác suất dự đoán)
Tuning	C, loss, calibration , cross-validation	n_estimators, max_depth, min_samples, class_weight, ccp_alpha	C, penalty, solver, class_weight, max_iter	Params, StratifiedKFold, Cross-Validation

MÔ HÌNH VÀ PHƯƠNG PHÁP



So sánh các mô hình phân lớp



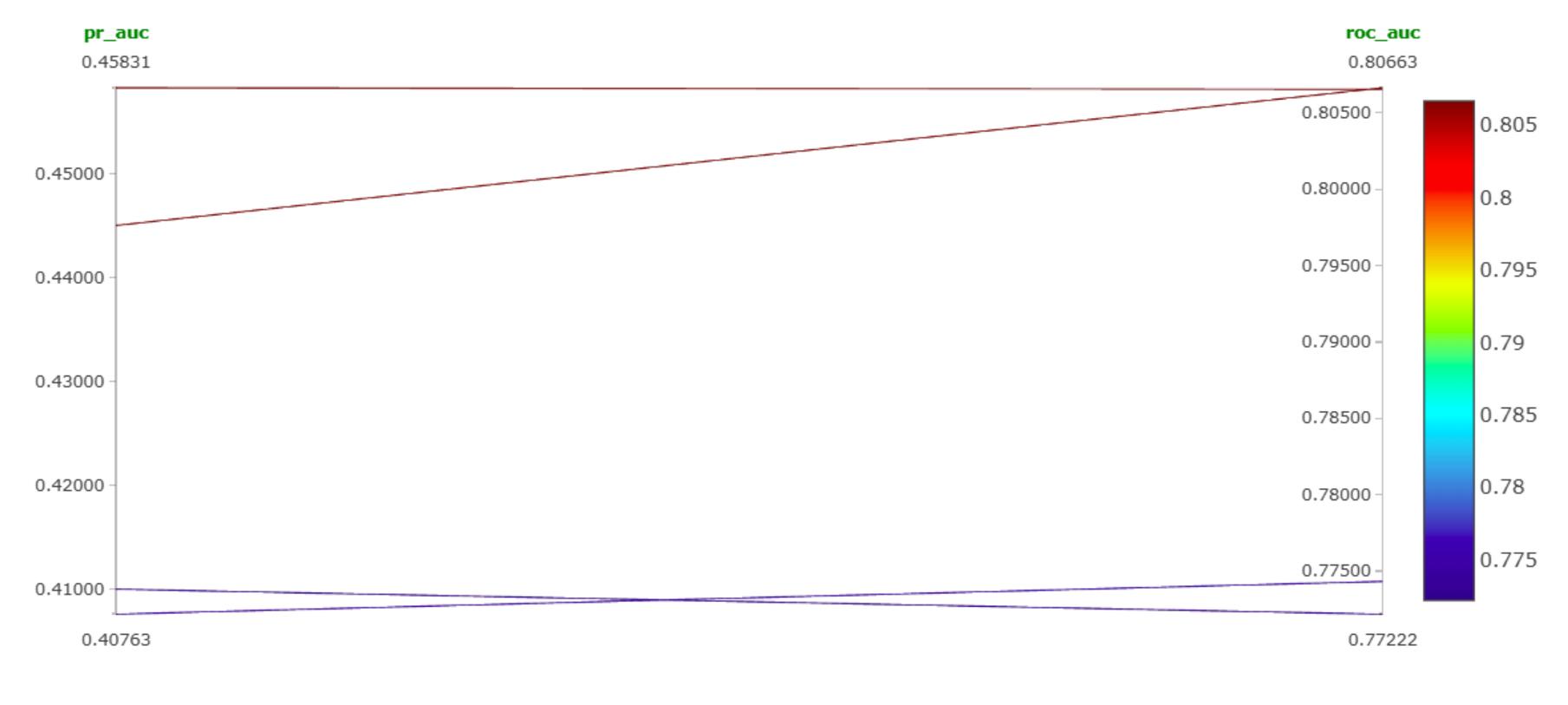
Recall cao - Precision thấp

- Logistic Regression
- Random Forest

Recall thấp- Precision cao

- SVM
- Gradient Boosting

So sánh các mô hình phân lớp



PR-AUC và ROC-AUC cao

- Gradient Boosting
- Random Forest

PR-AUC và ROC-AUC thấp

- SVM
- Logistic Regresion

Chỉ số	Ý nghĩa	SVM	Random Forest	Gradient Boosting	Logistic Regression
Recall cao	Tìm được nhiều KH tiềm năng		x		x
	Nhiều khả năng gọi nhầm KH không có nhu cầu				
Precision cao	Dự đoán tốt khách hàng sẽ đăng ký	x		x	
	Nhiều khả năng bỏ sót KH tiềm năng				
ROC-AUC cao	Phân biệt tốt tiềm năng khách hàng		x	x	
PR-AUC cao	Chọn khách hàng tiềm năng hiệu quả		x	x	

MODEL PERFORMANCE EVALUATION

	SVM	Random Forest	Logistic Regression	Gradient Boosting
Accuracy	0.89	0.85	0.76	0.89
Precision	0.64	0.41	0.27	0.61
Recall	0.19	0.57	0.65	0.26
F1	0.3	0.48	0.39	0.36
ROC_AUC	0.77	0.81	0.77	0.81
PR_AUC	0.41	0.45	0.41	0.46

BUSINESS PERFORMANCE EVALUATION

Conversion Rate khi ưu tiên gọi Top 15% khách hàng theo xếp hạng P(yes)

	SVM	Random Forest	Logistic Regression	Gradient Boosting
Conversion Rate	39%	43%	38%	44%
Conversion Rate ban đầu: ~11%				

Phân tích lỗi – Gradient Boosting

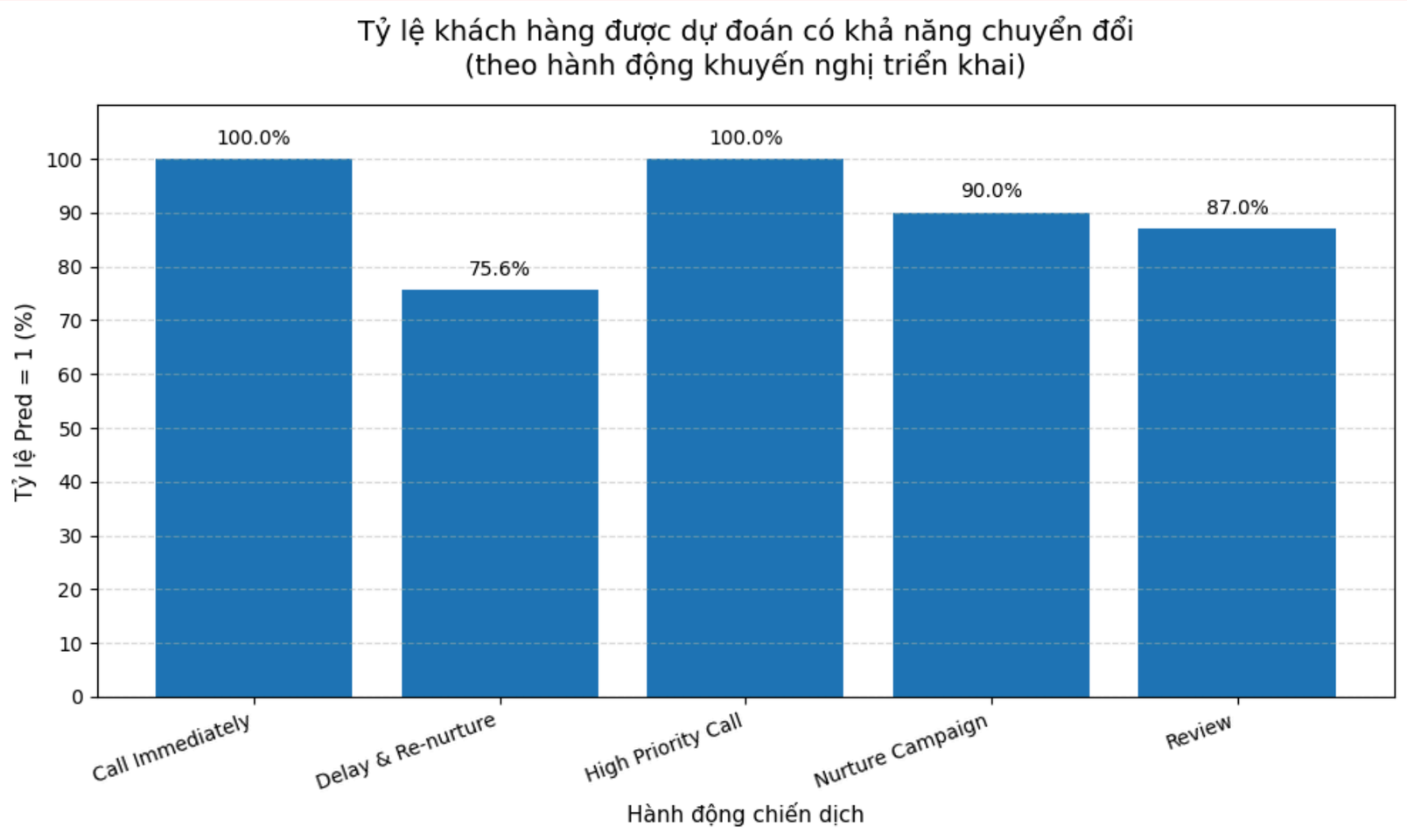
False Positive - Sai lầm loại 1 (gọi lãng phí)

- Lịch sử chiến dịch tốt trong quá khứ (poutcome=success, previous>0)
- Số dư cao, học vấn cao (management, tertiary)
- Chủ yếu qua cellular
- Hay xuất hiện ở tháng 4, tháng 6, tháng 10

False Negative - Sai lầm loại 2 (bỏ lỡ khách tiềm năng)

- Khách mới, không có lịch sử (previous=0, poutcome=no_previous_campaign)
- Số dư ngân hàng thấp/âm
- Nghề & học vấn thấp (blue-collar, services, retired; primary/secondary)
- Thường rơi vào tháng 5 - tháng 7

ỨNG DỤNG TRI THỨC VÀ MÔ HÌNH



Những khách hàng được khuyến nghị gọi ngay hoặc xếp vào nhóm ưu tiên cao gần như đều có khả năng đăng ký thành công theo dự đoán của mô hình.

→ Cho thấy tri thức từ việc khai phá dữ liệu là đúng

KẾT QUẢ ĐẠT ĐƯỢC VÀ HƯỚNG PHÁT TRIỂN

01 Kết quả đạt được

- Yếu tố quyết định: Kết quả chiến dịch trước (poutcome) ảnh hưởng mạnh nhất; Vay mua nhà (housing) là rào cản lớn nhất.
- Hiệu quả vận hành: Xác định ngưỡng thời lượng gọi (120s) và điểm bão hòa số lần gọi.

02 Hạn chế

- Dữ liệu: Mất cân bằng dữ liệu (Imbalanced data).
- Thuộc tính: Thiếu dữ liệu hành vi sâu (lịch sử giao dịch chi tiết, tương tác đa kênh).

03 Hướng phát triển

- Mở rộng dữ liệu: Đa dạng hóa nguồn dữ liệu ngân hàng và thời gian.
- Bổ sung đặc trưng: Tích hợp lịch sử giao dịch, chỉ số gắn kết khách hàng.
- Nâng cấp kỹ thuật: Áp dụng Deep Learning để khai thác các mối quan hệ phi tuyến phức tạp.

THANK YOU !