

## TÌNH HUỐNG

**GM Finance** là công ty tài chính tiêu dùng với hơn 20 năm hoạt động tại Việt Nam, cung cấp giải pháp tín dụng nhanh chóng- linh hoạt - tin cậy.

## THỬ THÁCH

Phân bổ lead không phù hợp với kinh nghiệm của Sales Agent (SA)

Tỉ lệ chuyển đổi lead ở các lead source chênh lệch lớn

Công ty muốn **tăng doanh thu và hiệu suất bán hàng** thông qua hệ thống **LDS** (Lead Distribution System)

Không đủ lead mỗi ngày cho mỗi SA để khai thác

Offer không đúng sản phẩm thực sự khách hàng muốn mua

## CÂU HỎI

Làm thế nào để cải thiện hệ thống LDS bằng cách sử dụng mô hình ML, giúp: Phân bổ lead phù hợp hơn với SA. Tối ưu hóa việc phân khúc khách hàng và chọn đúng sản phẩm phù hợp. Rút ngắn thời gian bán hàng và nâng cao hiệu quả trên mỗi SA?

## GIẢI PHÁP

### Machine Learning Methods:

- Model 1:** Mô hình dự đoán khách hàng có tiềm năng được duyệt hợp đồng
- Model 2:** Mô hình gán lead tối ưu cho SA dựa trên hiệu suất quá khứ, kinh nghiệm.

### Chiến lược triển khai “Smart Leads Distribution System” trong 1 năm

## TÁC ĐỘNG

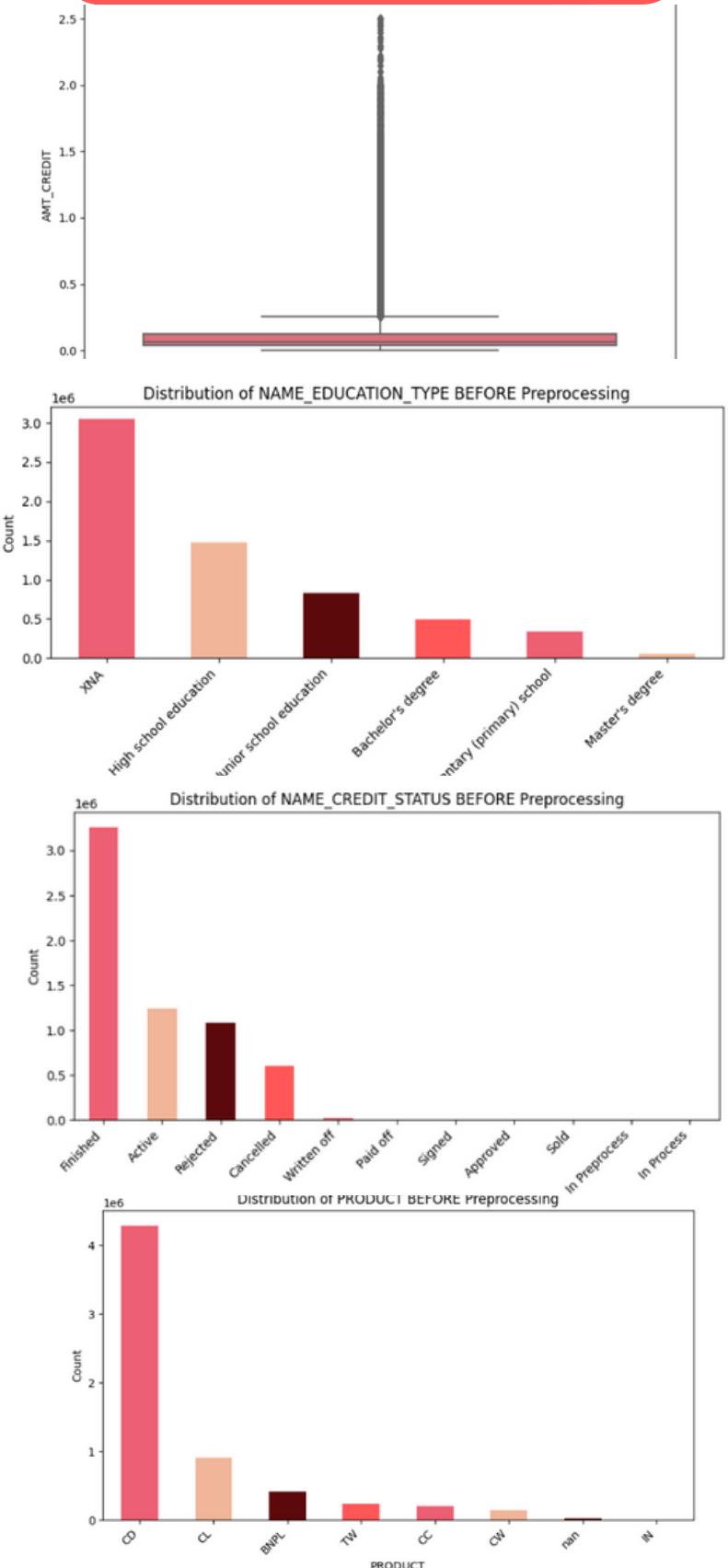
+33% Cải thiện tỷ lệ chuyển đổi lead

+57% Tăng hiệu suất trên mỗi SA

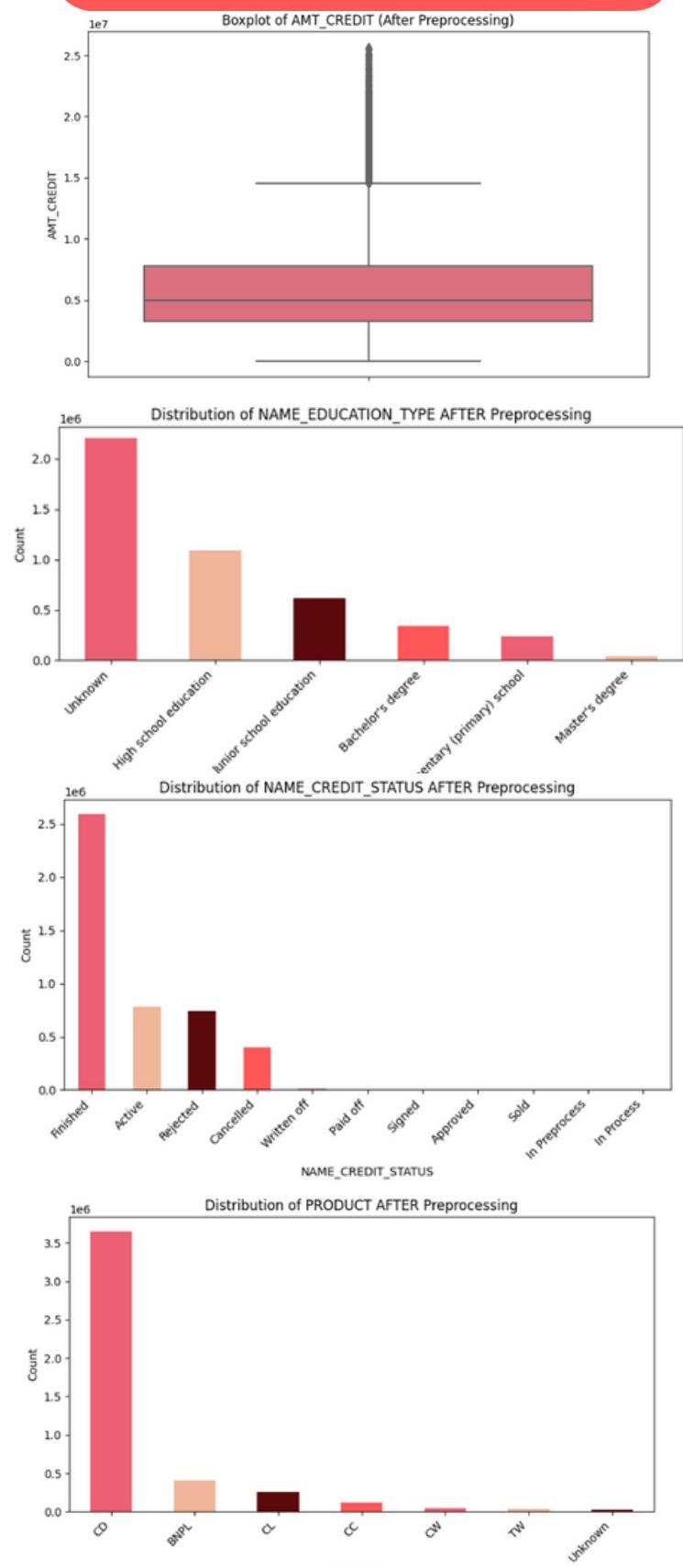
# TIỀN XỬ LÍ DỮ LIỆU - DATA PREPROCESSING

Tình trạng	Tiêu chí phân loại	Phương pháp xử lý
<b>Outlier</b>	Dựa theo IQR (Q1, Q3, IQR = Q3 - Q1)	<ul style="list-style-type: none"> <li>Xác định outlier: ngoài <math>[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]</math></li> <li>Xử lý: điền bằng median</li> </ul>
<b>Null values – Numerical</b>	Tỷ lệ thiếu < 20%	Điền bằng median
	Tỷ lệ thiếu $\geq 20\%$	Gán giá trị là "Unknown" hoặc tạo cờ nhị phân (missing flag)
<b>Null values – Categorical</b>	null, 'XNA', rỗng	Gán đồng loạt là "Unknown"
<b>Không đóng góp nhiều trong quá trình phân tích</b>	tỉ lệ null > 90%, ID	drop

## Trước Preprocessing

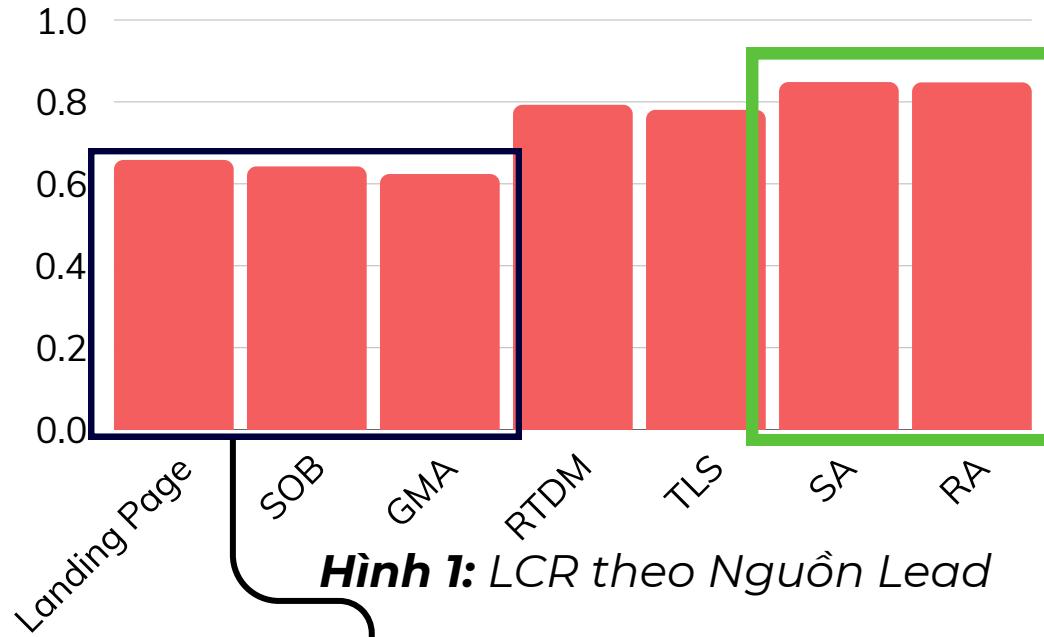


## Sau Preprocessing



# THÁCH THỨC TỪ CHÊNH LỆCH TỶ LỆ CHUYỂN ĐỔI THEO NGUỒN LEAD & QUY TRÌNH

Tỉ lệ chuyển đổi Leads thấp: **35.25 %**



Hình 1: LCR theo Nguồn Lead



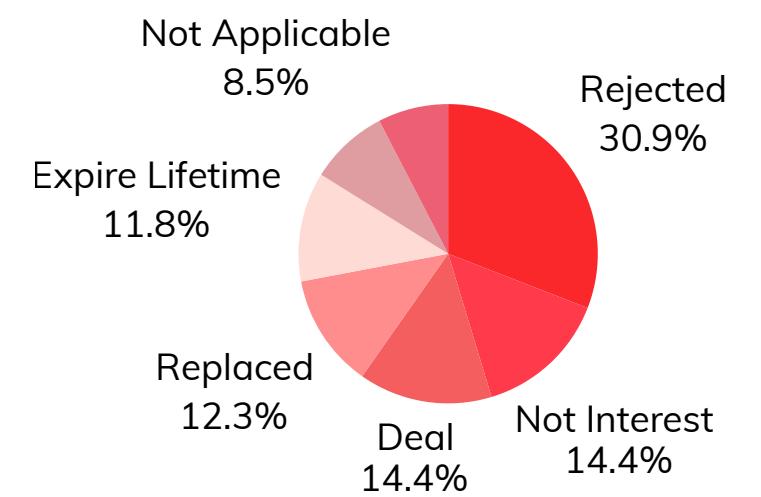
Hình 2: LCR theo Quy trình



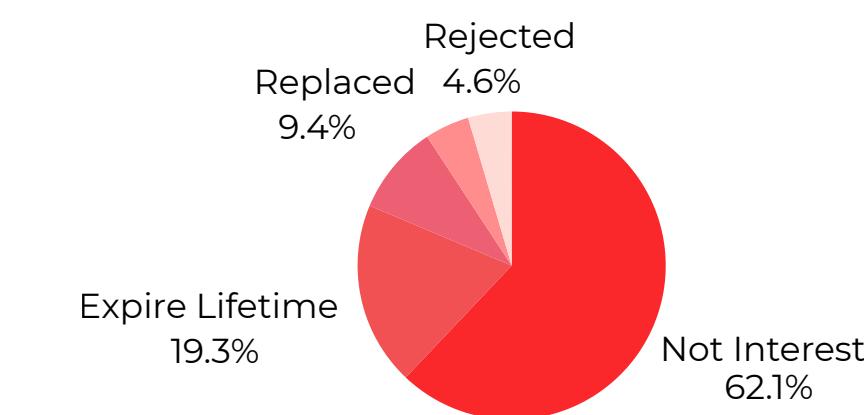
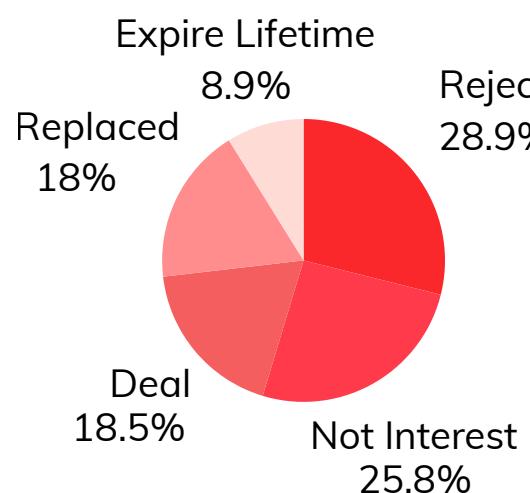
Tỉ lệ chuyển đổi cao nhờ:  
**tư vấn trực tiếp + kiểm soát quy trình tốt**

Điều này cho thấy vai trò quan trọng của việc đầu tư vào đội ngũ nội bộ và kiểm soát tốt trải nghiệm khách hàng đầu cuối.

## Cụ thể hơn lí do phê duyệt thấp



Hình 3: Tỉ lệ Rejected Leads theo SOB & Landing



Hình 4: Tỉ lệ Rejected Leads theo GMA

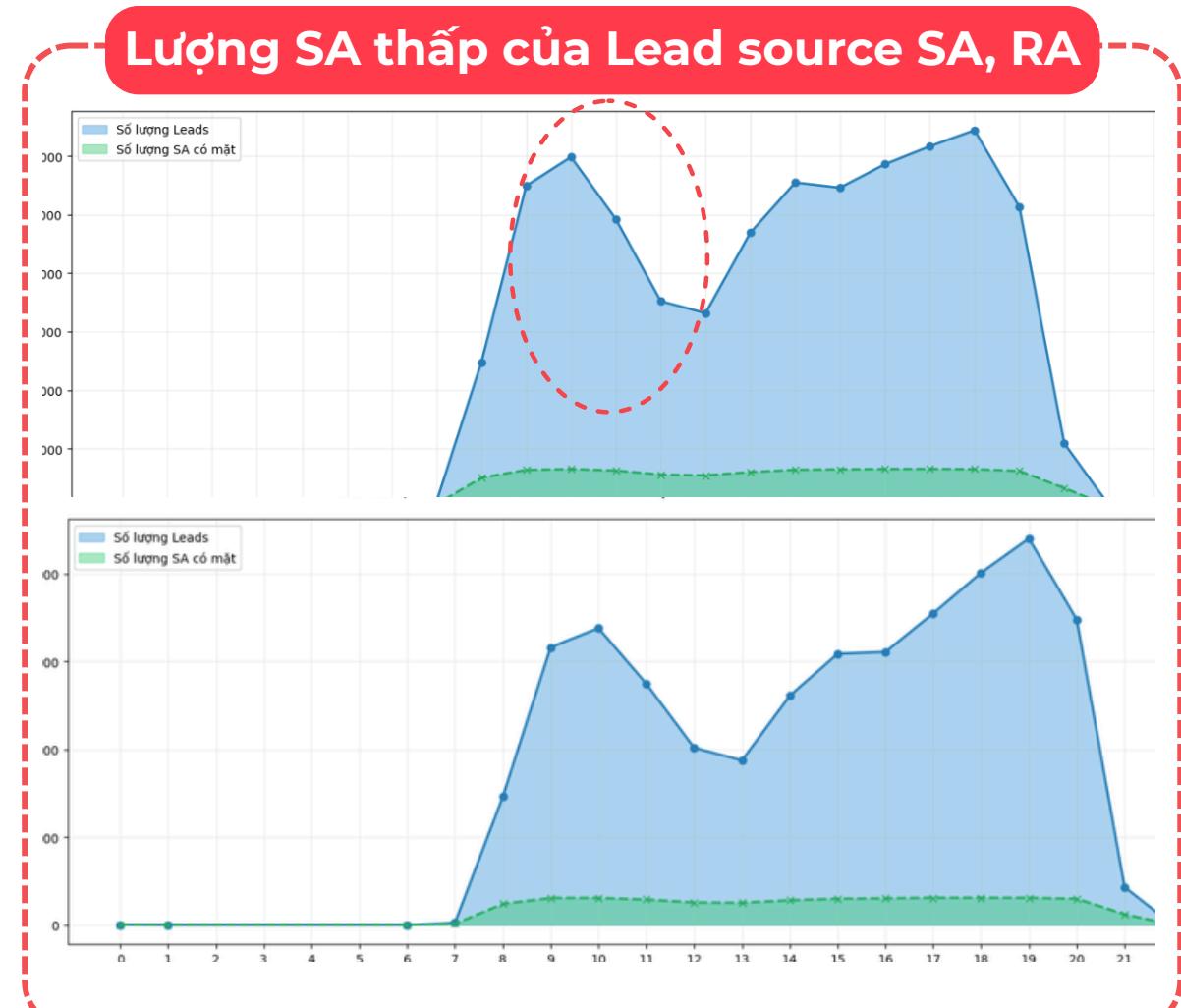
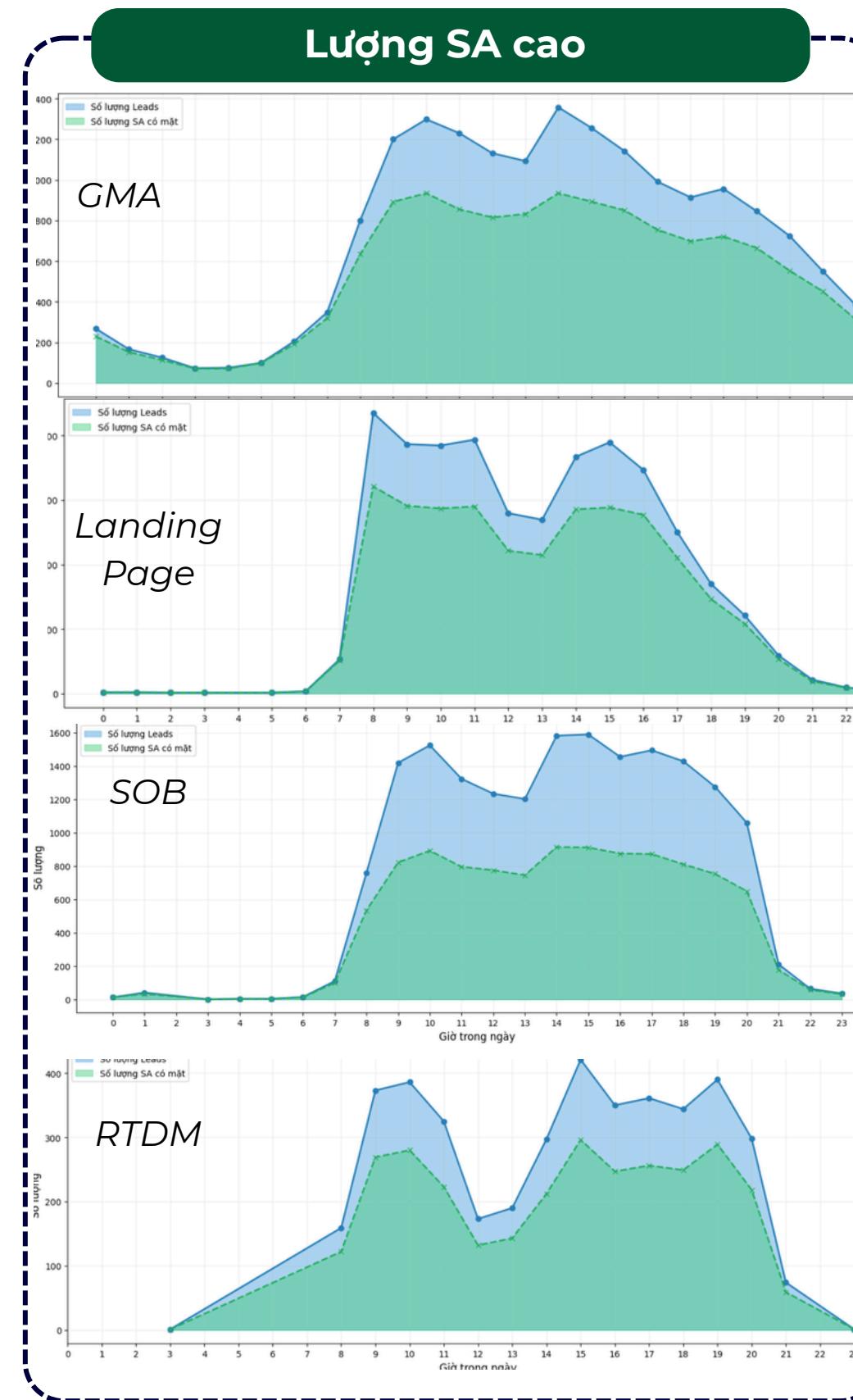
Nhầm tệp khách hàng hoặc **không phù hợp** với quy trình hiện tại.

💡 Phát triển hệ thống nhận diện khách hàng **tiềm năng** **được duyệt**

Nguồn khách hàng **chất lượng thấp**, có tiềm năng nếu cải thiện quy trình chăm sóc

💡 Phát triển hệ thống **phân khúc khách hàng** để nhận diện khách hàng chất lượng cao.

# TÌNH TRẠNG PHÂN BỐ SA CHƯA TỐI ƯU THEO HIỆU SUẤT NGUỒN LEAD



Hình 5: Phân phối SA theo Lead Sources

Số lượng Leads  
Số lượng SA có mặt

Lead Source	Conversion Rate (CR)	Đặc điểm
SA	Rất cao	Leads dồn giờ cao điểm, SA thiếu → dễ mất cơ hội
RA	Rất cao	Giống SA
RTDM	Trung bình	Số leads vừa phải, SA khớp
TLS	Trung bình	Phân phối SA hợp lý
GMA	Thấp	SA nhiều nhưng CR thấp → lãng phí giờ trưa-chiều
Landing Page	Thấp	Dồn nhiều leads vào buổi sáng
SOB	Thấp	Leads cao 10h-14h, SA không đủ → quá tải nhẹ

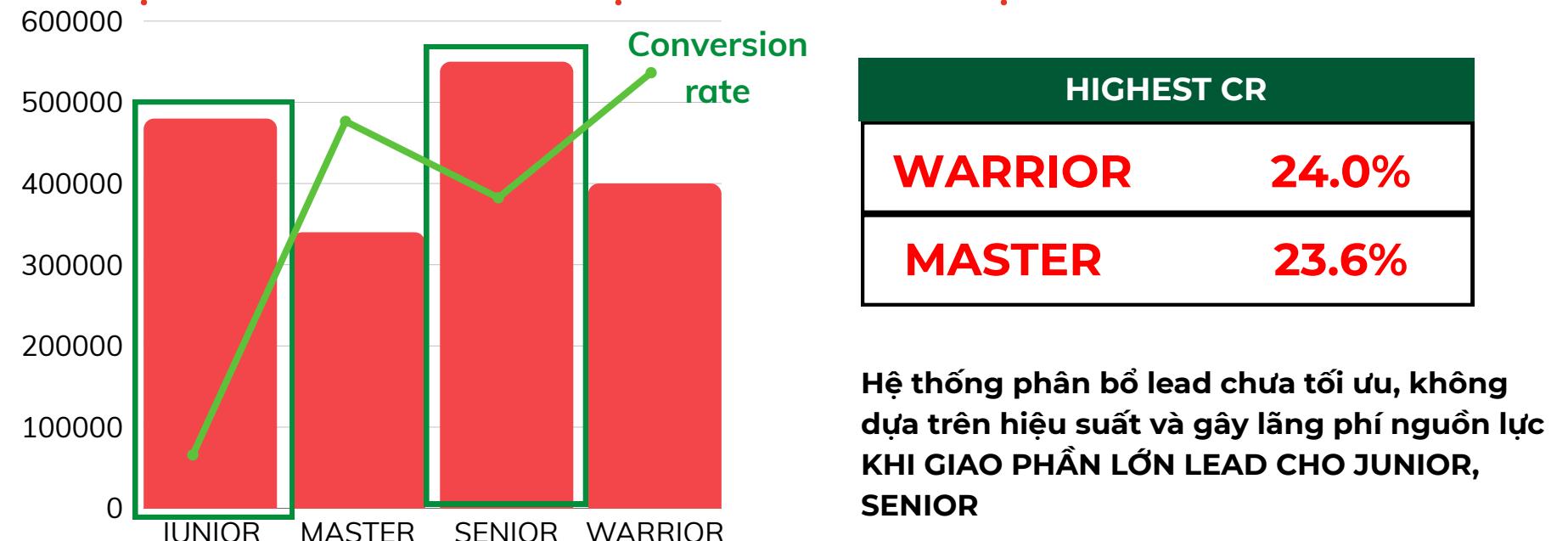
### Giải pháp

### Đề xuất phân bổ SA theo khung giờ

Nguồn Giảm SA (CR thấp)	Nguồn Ưu tiên (Tăng SA)
08h-11h	SOB, Landing Page → <b>SA, RA</b>
13h-15h	GMA → <b>RA</b>
17h-20h	SOB, Landing Page → <b>SA, RA</b>

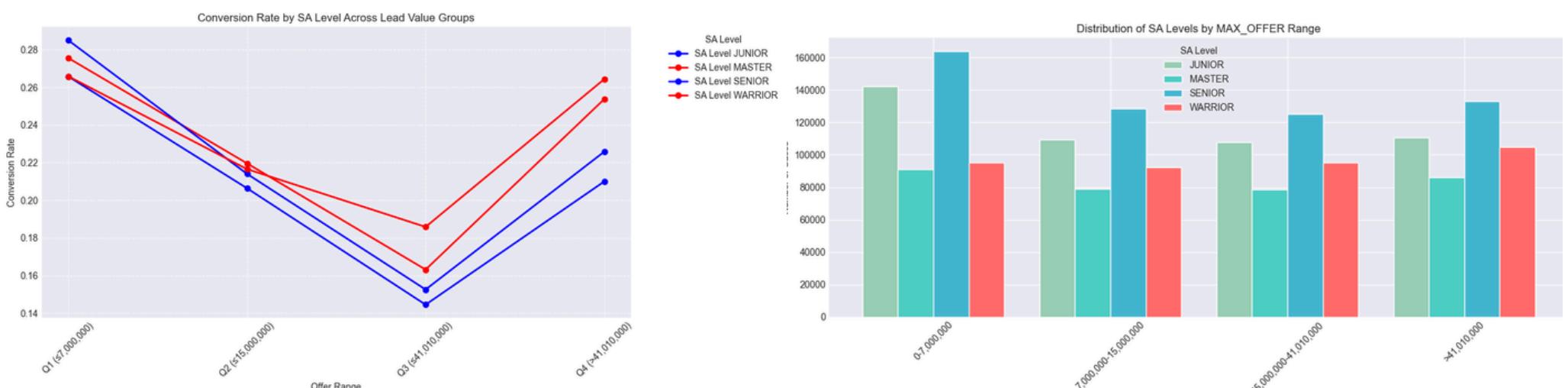
# THỰC TRẠNG PHÂN BỐ SALE ASSISTANT KÉM HIỆU QUẢ

## A. BẤT HỢP LÝ GIỮA KHỐI LƯỢNG LEAD VÀ HIỆU SUẤT SA



Hình 6: Tổng Leads và Conversion Rate theo SA level

## B. CHƯA PHÂN LOẠI ĐÚNG SA VỚI ĐỘ NÓNG CỦA LEAD



Tỉ lệ chuyển đổi của các case lead nóng, giá trị của max\_offer cao (từ trên 41 triệu) đối với các SA có trình độ cao (Warrior & Master) có khoảng cách tăng rõ rệt với các SA có trình độ thấp hơn. Tuy nhiên, hệ thống hiện tại lại đang phân bổ nhiều case lead nóng cho sa low level hơn.

### SOLUTION A

Thay đổi hệ thống phân phối, ưu tiên phân phối nhiều lead hơn nhóm SA có level cao như warrior và master.  
Đồng thời giảm tải cho các sa có level thấp hơn, điều này cho phép họ có nhiều THỜI gian chăm sóc lead hơn → **tăng tỉ lệ chuyển đổi lên**.

### KEY INSIGHT

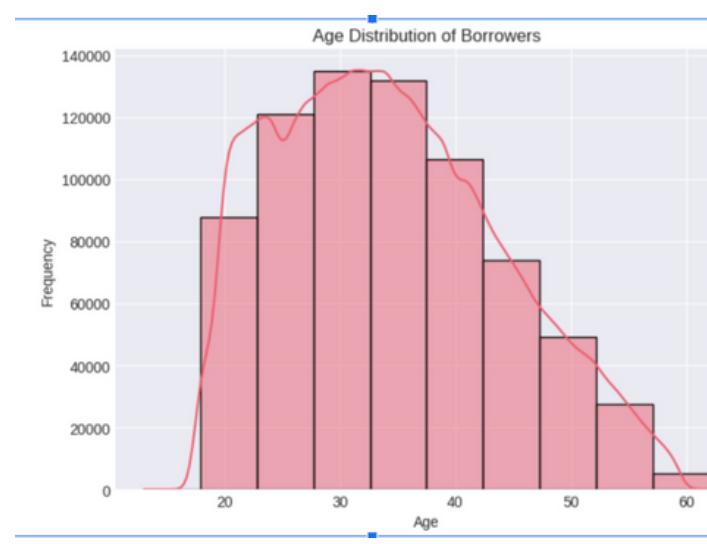
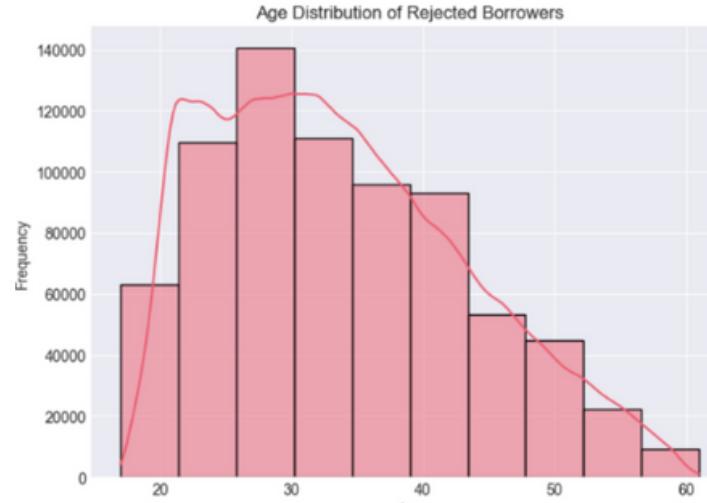
**HỆ THỐNG CHƯA TẬN DỤNG HIỆU QUẢ TRÌNH ĐỘ VÀ KINH NGHIỆM CỦA CÁC SA CẤP CAO.**

### SOLUTION B

Ưu tiên giao các **lead nóng**, các lead có giá trị cao cho các **SA có level Warrior và Master** bởi vì các lead có giá trị tài chính lớn, phức tạp hơn đòi hỏi kỹ năng tư vấn, đàm phán và xử lý từ chối ở một cấp độ cao hơn.

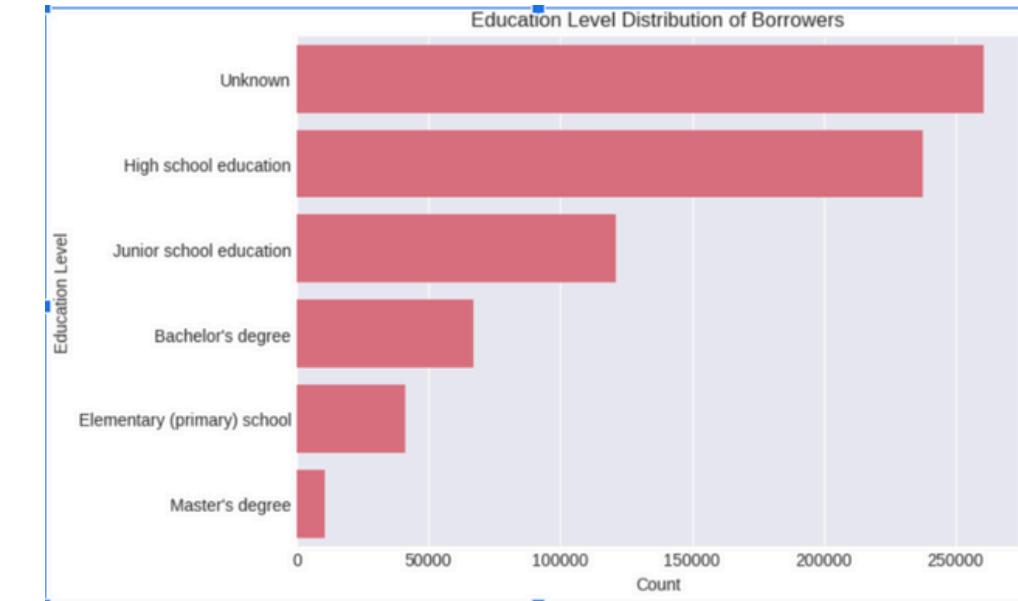
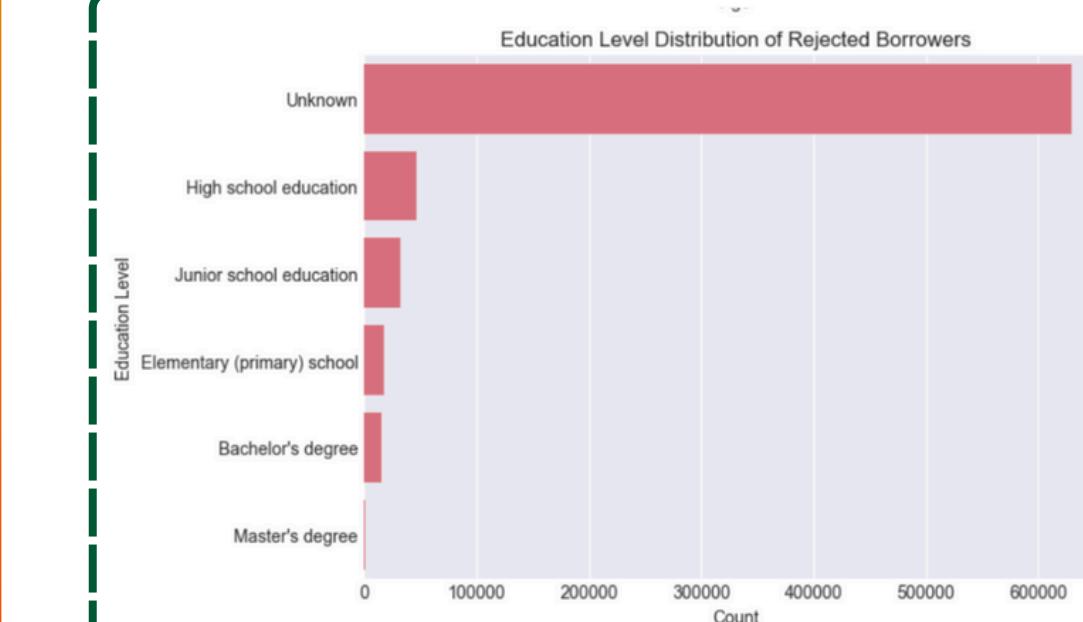
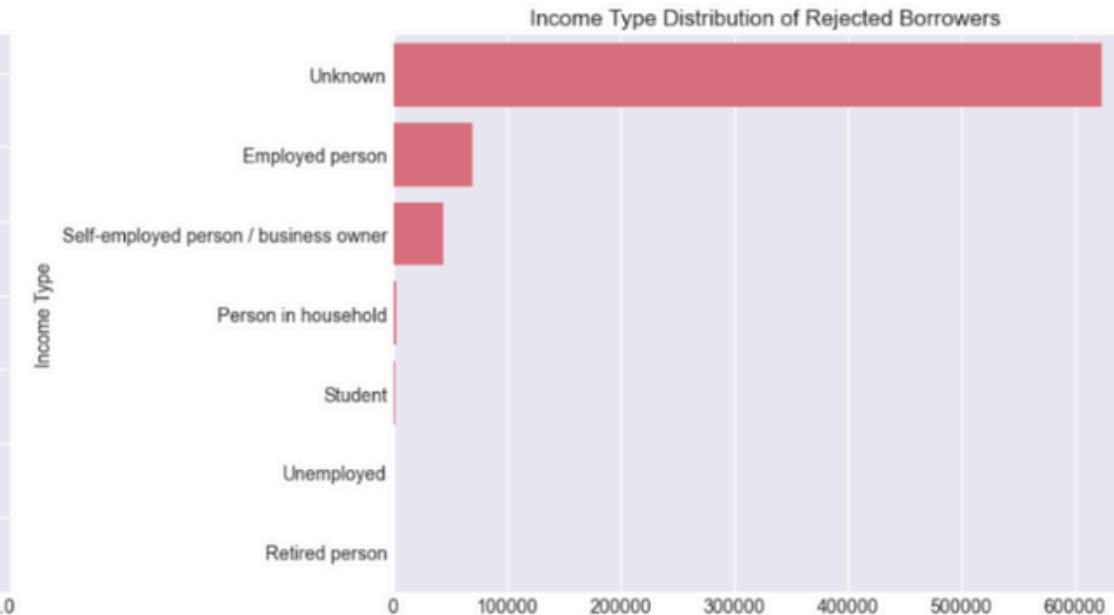
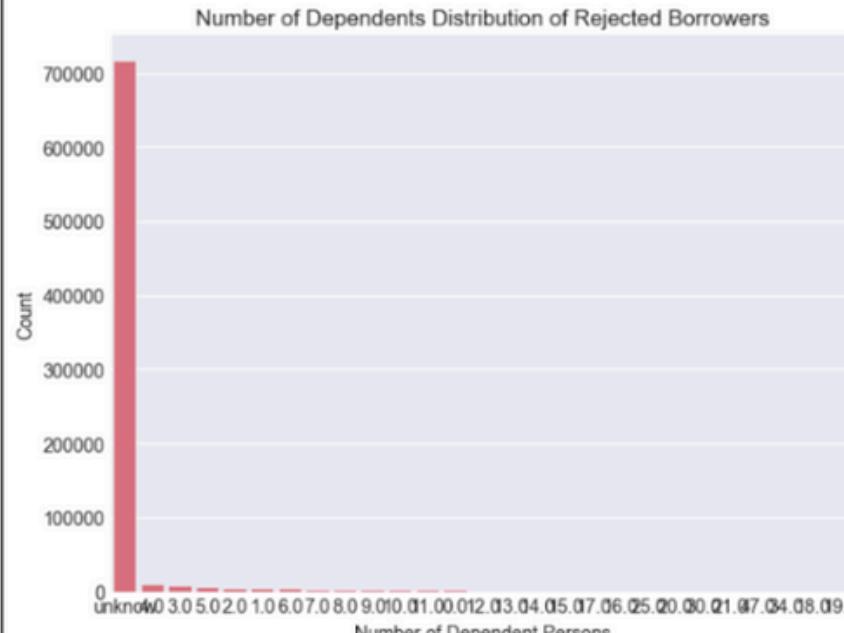
Ưu tiên các lead dễ, các lead giá trị thấp cho các SA ở level thấp hơn để họ trau dồi và học hỏi kinh nghiệm

# PHÂN TÍCH NHÂN KHẨU HỌC CỦA KHÁCH HÀNG



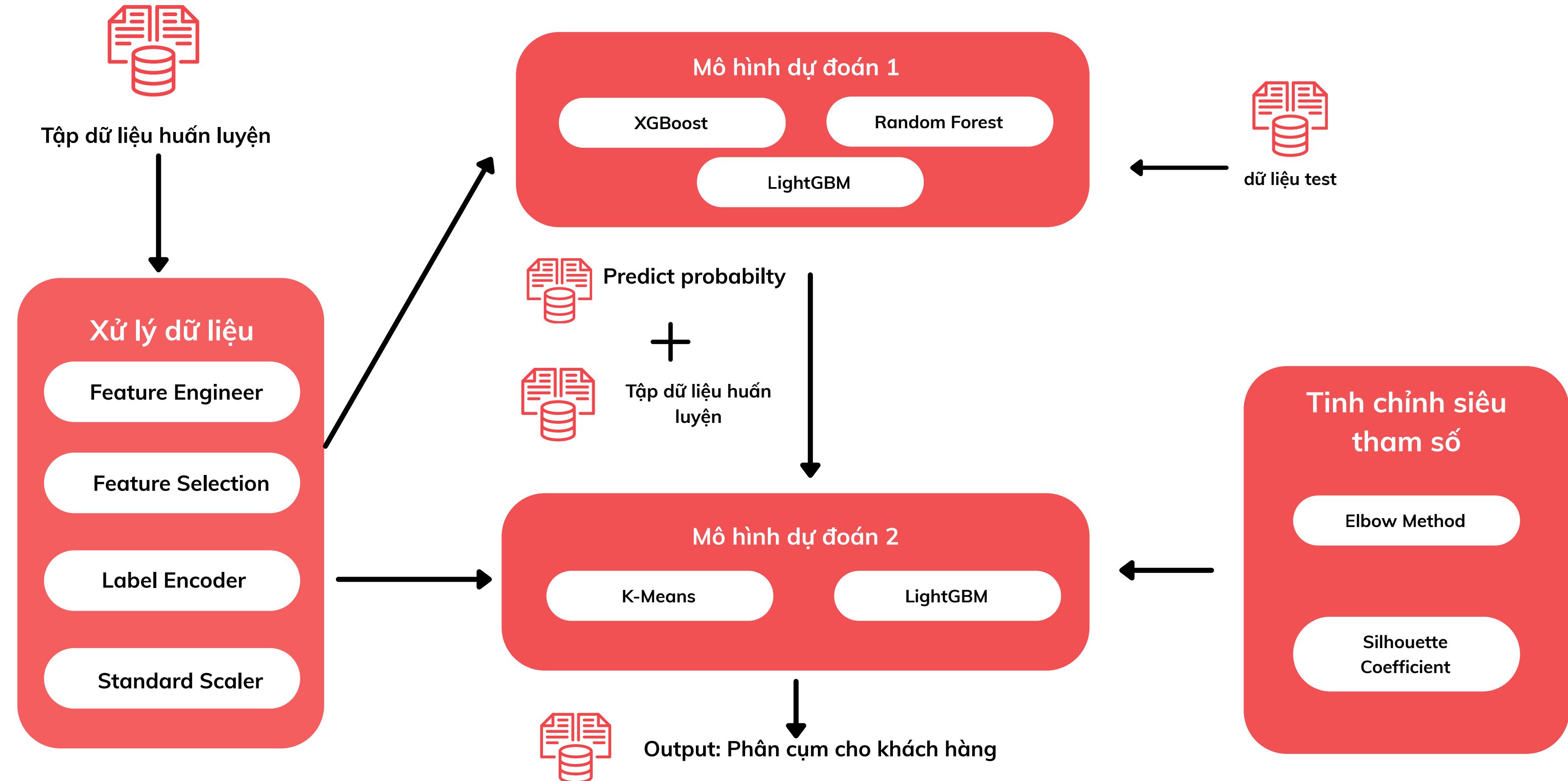
- Qua quan sát, có thể thấy nhân khẩu học của những khách hàng bị từ chối và được duyệt hợp đồng không quá khác nhau.
- Tuy nhiên điểm khác biệt lớn nhất là những khách hàng bị từ chối có nhiều biến unknown hơn.

**Vì vậy có thể triển khai chính sách mới hỗ trợ những khách hàng bị thiếu thông tin, đưa nhân viên hỗ trợ khách hàng điền thông tin đầy đủ hơn để có thể tăng tỉ lệ được cho vay**



- Khách hàng bị từ chối có thông tin về học vấn bị thiếu chiếm một số lượng rất lớn trong data.
- Trong khi đó ở chiều ngược lại, tỉ lệ thông tin bị thiếu ở những khách hàng được duyệt hợp đồng chỉ chiếm dưới 50%

# Huấn luyện mô hình



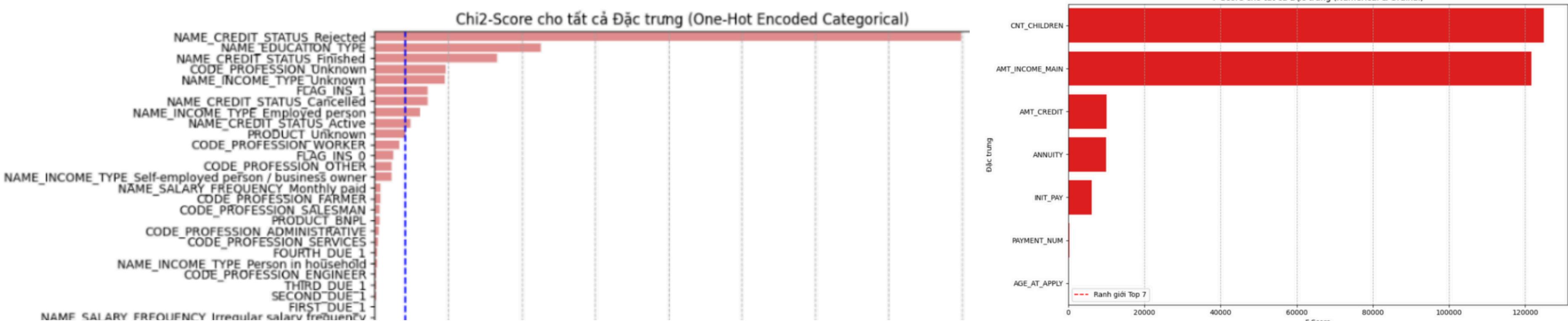
# Huấn luyện mô hình 1

Dùng Random Forest  
(55 đặc trưng)

Accuracy	0.9998
AUC-ROC	1.0000

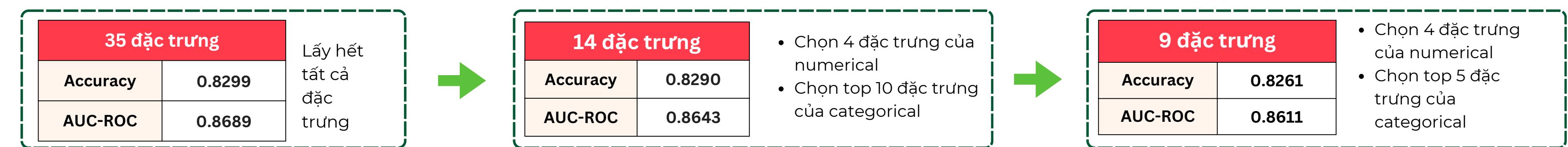
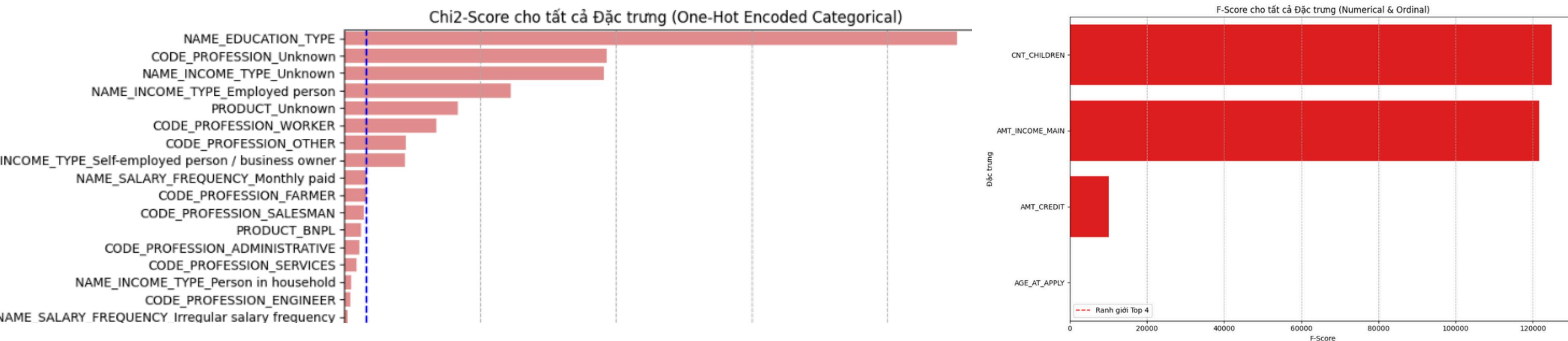
Điểm số kiểm tra cao một cách không thực tế:

- Mô hình đang nhìn thấy thông tin mà nó không nên có trong quá trình huấn luyện
- Dấu hiệu của rò rỉ dữ liệu (Data Leakage)



Loại bỏ các biến bị tình nghi là rò rỉ dữ liệu 'INIT\_PAY', 'ANNUITY', 'PAYMENT\_NUM', 'FIRST\_DUE', 'SECOND\_DUE', 'THIRD\_DUE', 'FOURTH\_DUE', 'FLAG\_INS', 'NAME\_CREDIT\_STATUS'

# Loại bỏ biến gây rò rỉ dữ liệu - Chọn đặc trưng tối ưu nhất



Lựa chọn 14 đặc trưng là  
số đặc trưng tối ưu

- Duy trì thông tin quan trọng
- Giảm thiểu độ phức tạp của mô hình
- Khác biệt là không đáng kể

## Mục tiêu: mô hình dự đoán khách hàng tiềm năng “Approved” hoặc “Not Approved”

- Mô hình LightGBM có điểm AUC-ROC cao nhất là **0.8899**
- Accuracy cao nhất là **0.8461**



**Chọn mô hình LightGBM để tinh chỉnh và đưa ra dự đoán**

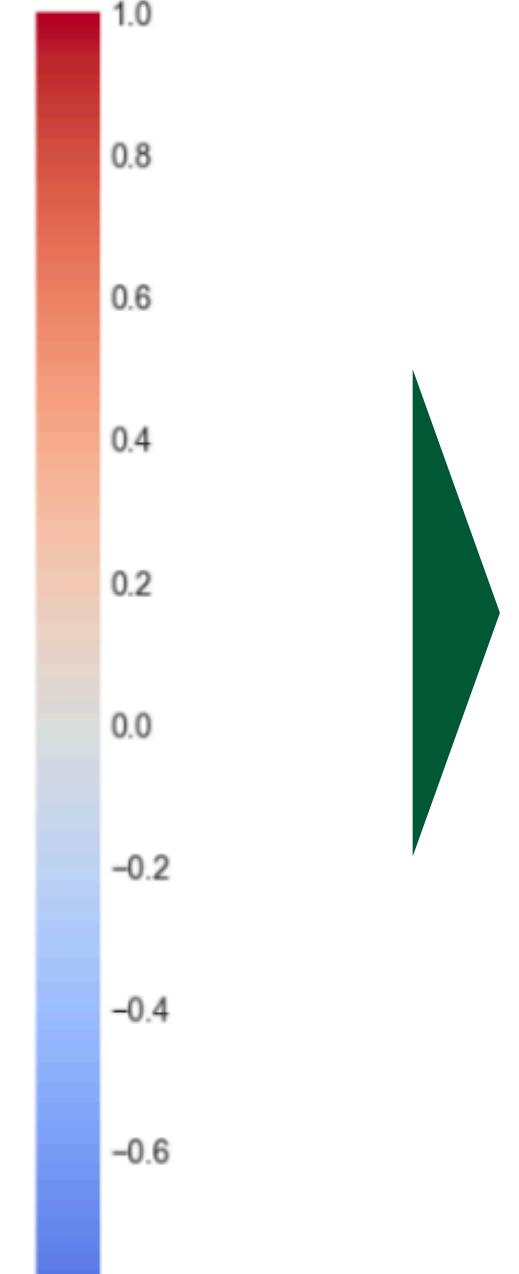
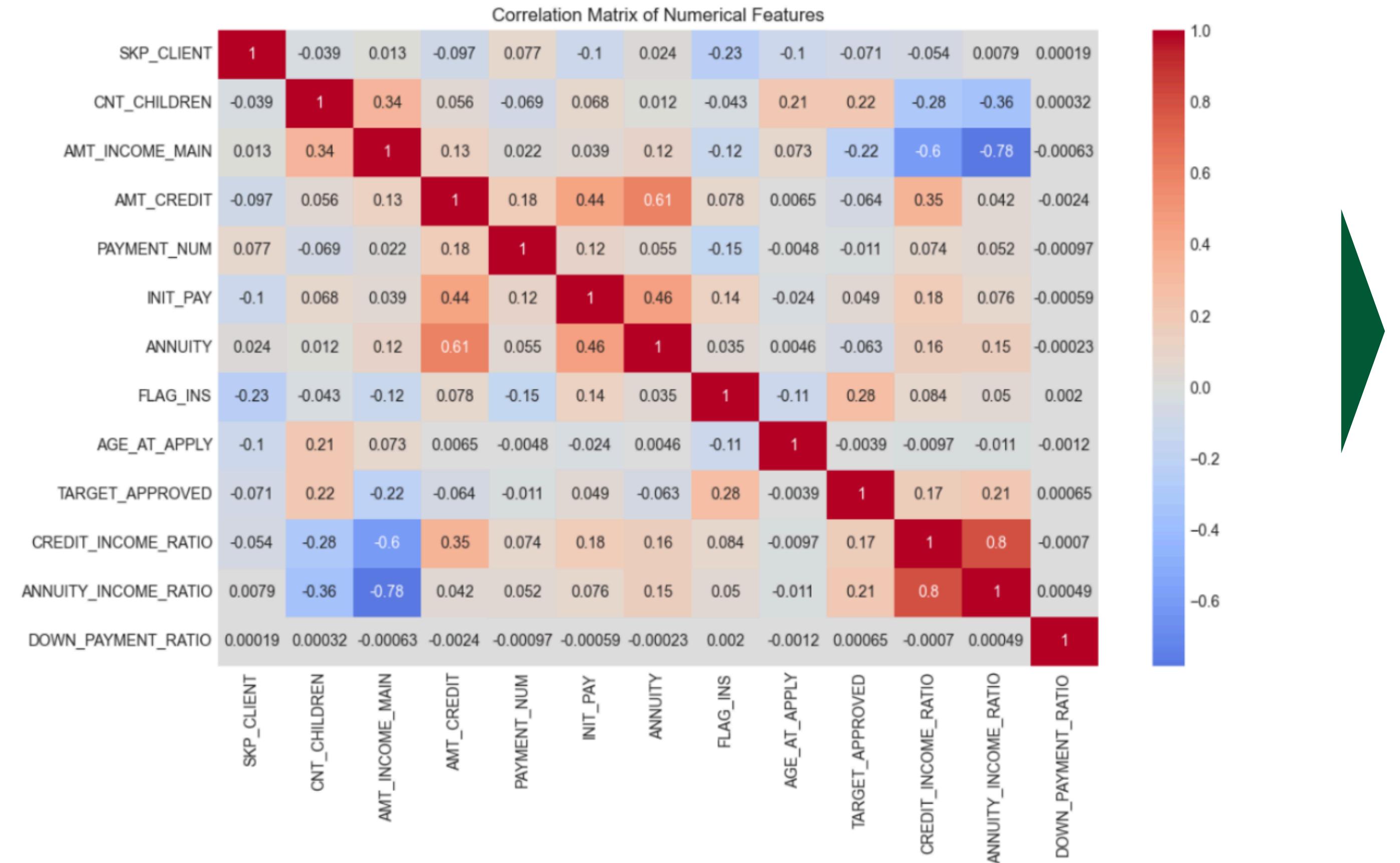
Tinh chỉnh mô hình bằng phương pháp tối ưu hóa Bayesian (Sử dụng Optuna)

### Model Performance Metrics Summary

Metric	XGBoost	RandomForest	LightGBM
Accuracy	0.8235	0.8290	0.8461
Precision	0.8100	0.8500	0.8200
Recall	0.8100	0.8200	0.8100
F1-Score	0.8100	0.8300	0.8200
AUC-ROC	0.8888	0.8643	0.8899

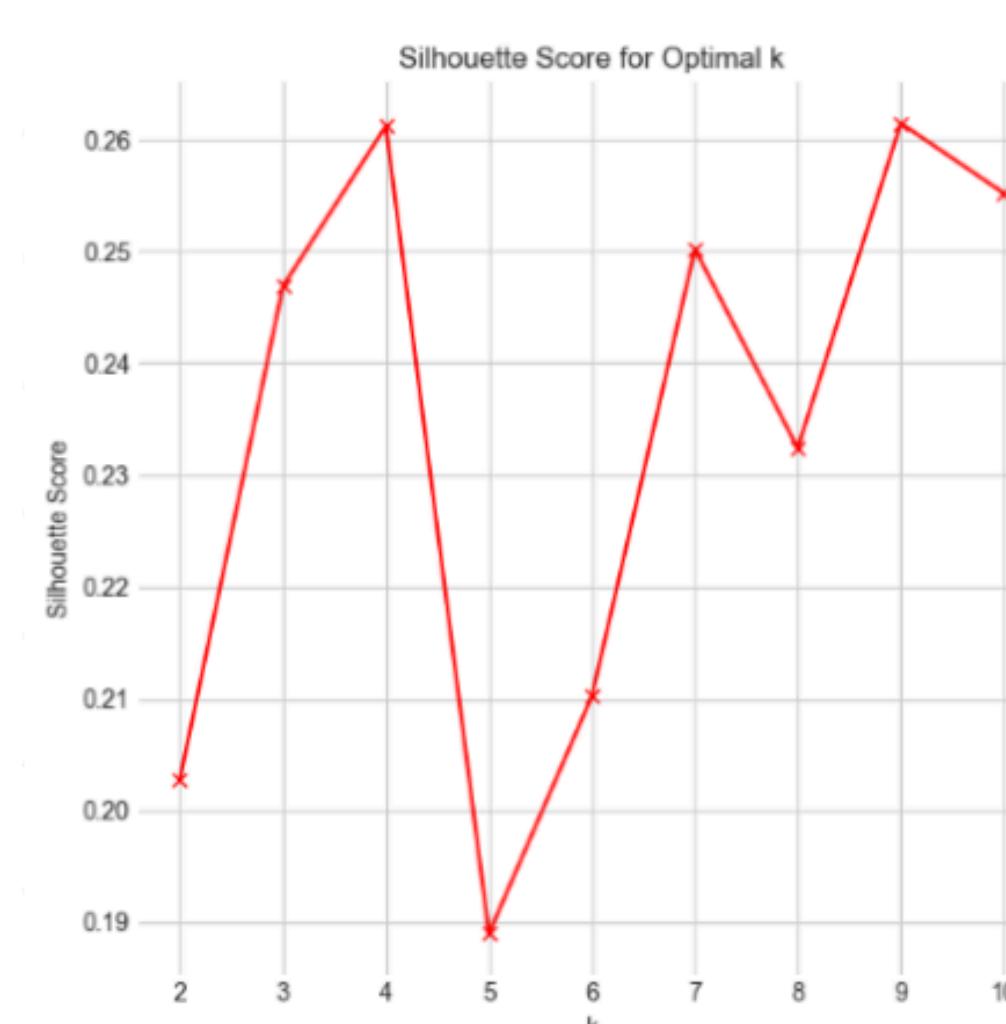
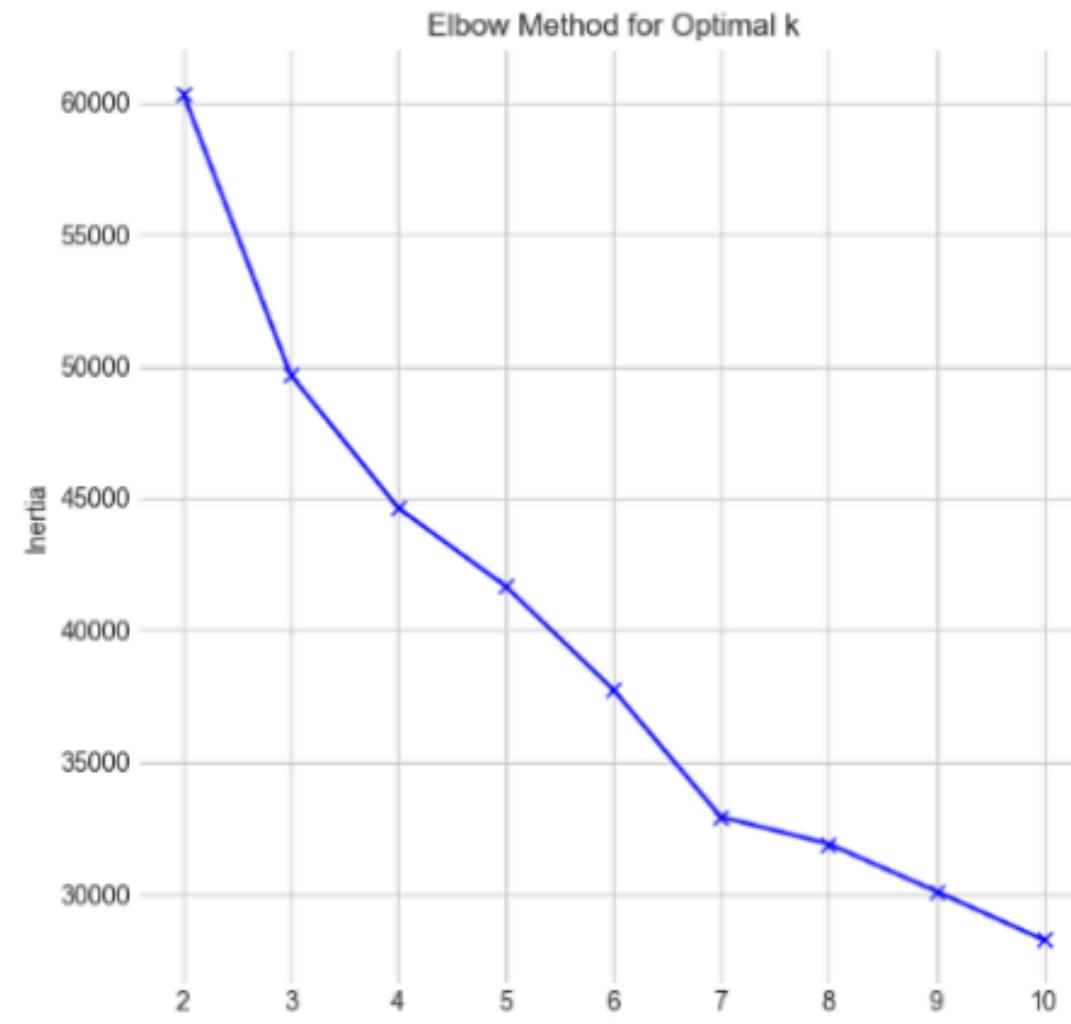
- AUC-ROC: 0.9100**
- Accuracy: 0.8561**

# Huấn luyện mô hình 2



Loại bỏ  
Credit\_Income\_Ratio vì  
nó gây ra vấn đề đe  
cộng tuyến 0.8

# Lựa Chọn Số Lượng Cụm Tối Ưu: Phân Tích Elbow và Silhouette



Dựa trên hai biểu đồ này, lựa chọn tốt nhất và đáng tin cậy nhất là k=4.

Phương pháp Elbow (Biểu đồ bên trái):

- Biểu đồ này cho thấy một "khuỷu tay" (elbow) rất rõ ràng tại k=4. Sau điểm này, độ dốc của đường cong giảm đi đáng kể, có nghĩa là việc thêm nhiều cụm hơn không mang lại hiệu quả cao trong việc giảm quán tính (Inertia).

Phương pháp Silhouette Score (Biểu đồ bên phải):

- Mặc dù điểm số cao nhất tuyệt đối là ở k=9, nhưng có một đỉnh rất cao và rõ ràng tại k=4.
- Điều quan trọng nhất là sau k=4, điểm số sụt giảm mạnh ở k=5. Điều này cho thấy cấu trúc 4 cụm là một cấu trúc tự nhiên và ổn định. Việc cố gắng thêm một cụm thứ 5 đã phá vỡ sự tách biệt rõ ràng này.

# Tổng quan 4 Phân khúc Khách hàng

## Cluster 1: Giá trị cao, Không cần thu nhập (High Value, Income Not Required)

### Đặc điểm chính

- Tỷ lệ duyệt cao nhất (**0.84**)
- Thông tin thu nhập không được cung cấp hoặc gần như bằng không (**AMT\_INCOME\_MAIN = 1.00**).
- Có tỷ lệ trả trước (DOWN\_PAYMENT\_RATIO) cao (**0.40**) - yếu tố then chốt

### Ý nghĩa kinh doanh

Đây là phân khúc khách hàng rất giá trị và có khả năng chuyển đổi cao. Họ có thể là chủ kinh doanh hoặc các khách hàng có tài sản đảm bảo khác, không dựa vào lương tháng.

### Chiến lược đề xuất

Ưu tiên phân bổ cho các SA cấp cao nhất (**Master, Warrior**). Các SA này có đủ kỹ năng để xử lý các hồ sơ phức tạp, không theo chuẩn thông thường và chốt các hợp đồng giá trị cao.

## Cluster 2: An toàn, Khoản vay nhỏ (Secure, Small Loans)

- Hồ sơ an toàn và đầy đủ thông tin nhất: có việc làm ("Employed person"), trình độ học vấn rõ ràng ("High school education").
- Thu nhập ở mức tốt (**trung bình 7.5M**) nhưng chỉ vay các khoản nhỏ (**trung bình 2.9M**).
- Tỷ lệ duyệt tốt (**0.76**).

Đây là nhóm khách hàng "dễ tính", rủi ro thấp và quy trình xử lý đơn giản.

Phân bổ cho các SA cấp **Junior** hoặc **Senior**. Điều này giúp xử lý nhanh chóng các hồ sơ đơn giản, đồng thời giải phóng thời gian của các SA cấp cao hơn cho những ca khó.

## Cluster 0: Thu nhập cao, Rủi ro tiềm ẩn (High Income, Hidden Risk)

### Đặc điểm chính

- Có thu nhập **(11.3M)** và khoản vay **(8.2M)** đều ở mức cao.
- Tình trạng thu nhập rõ ràng ("Employed person").
- Tuy nhiên, tỷ lệ duyệt chỉ ở mức **trung bình thấp (0.51)**.

### Ý nghĩa kinh doanh

Đây là nhóm khách hàng có giá trị cao nhưng khó chuyển đổi. Nguyên nhân có thể do họ "khó tính", có yêu cầu phức tạp, hoặc có các rủi ro ngầm khác (lịch sử tín dụng, thông tin không nhất quán).

### Chiến lược đề xuất

Yêu cầu kỹ năng xử lý phức tạp và khả năng thuyết phục cao. Cần được phân bổ cho các SA cấp **Warrior** hoặc **Master** để tối đa hóa cơ hội chuyển đổi, tránh lãng phí các lead giá trị.

## Cluster 3: Giao dịch Lỗi / Bất thường (Error / Anomalous Transactions)

- Tỷ lệ duyệt gần như bằng **không (0.04)**.
- Khoản vay (AMT\_CREDIT) cũng gần như **bằng không**.

Cụm này không đại diện cho một phân khúc khách hàng thực sự mà có khả năng là nơi tập trung các dữ liệu lỗi, hồ sơ test, hoặc các giao dịch không hợp lệ trong quá trình thu thập dữ liệu.

Lọc và loại bỏ khỏi quy trình phân bổ lead tự động. Các hồ sơ này nên được chuyển cho một bộ phận khác để kiểm tra và xác minh thủ công, tránh làm lãng phí thời gian và nguồn lực của đội ngũ bán hàng.

# TỔNG QUAN KẾ HOẠCH TRIỂN KHAI DỰ ÁN “SMART LEADS DISTRIBUTION SYSTEM”

KẾ HOẠCH



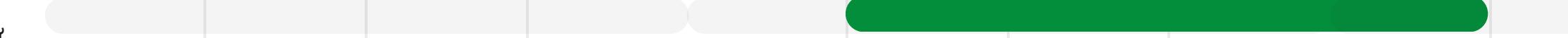
**Phase 1:** Làm sạch dữ liệu, chuẩn hóa hệ thống, xây dựng pipeline ML



**Phase 2:** Huấn luyện mô hình dự đoán khách hàng được kí, deploy LightGBM



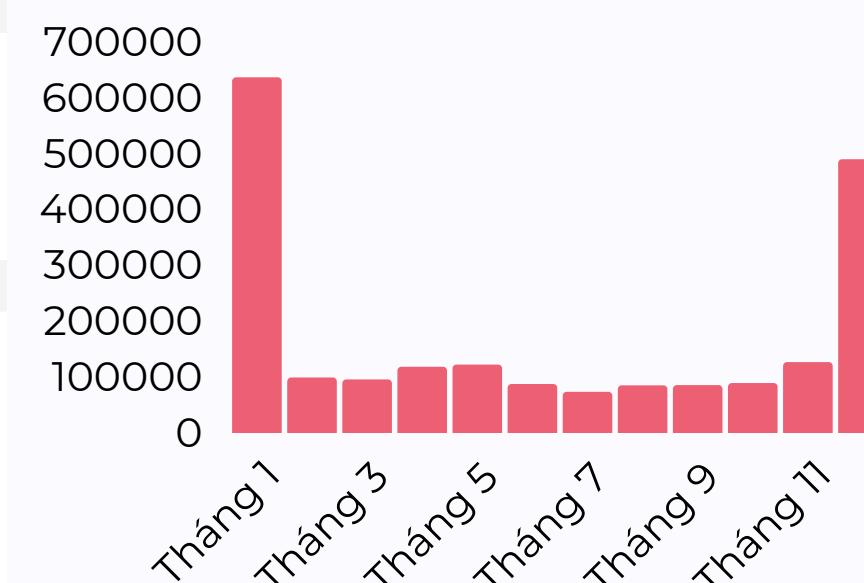
**Phase 3:** Phân cụm khách hàng bằng K-means, tạo nhóm xử lý ưu tiên



**Phase 4:** A/B testing LDS cũ vs mới, đo CR feedback SA

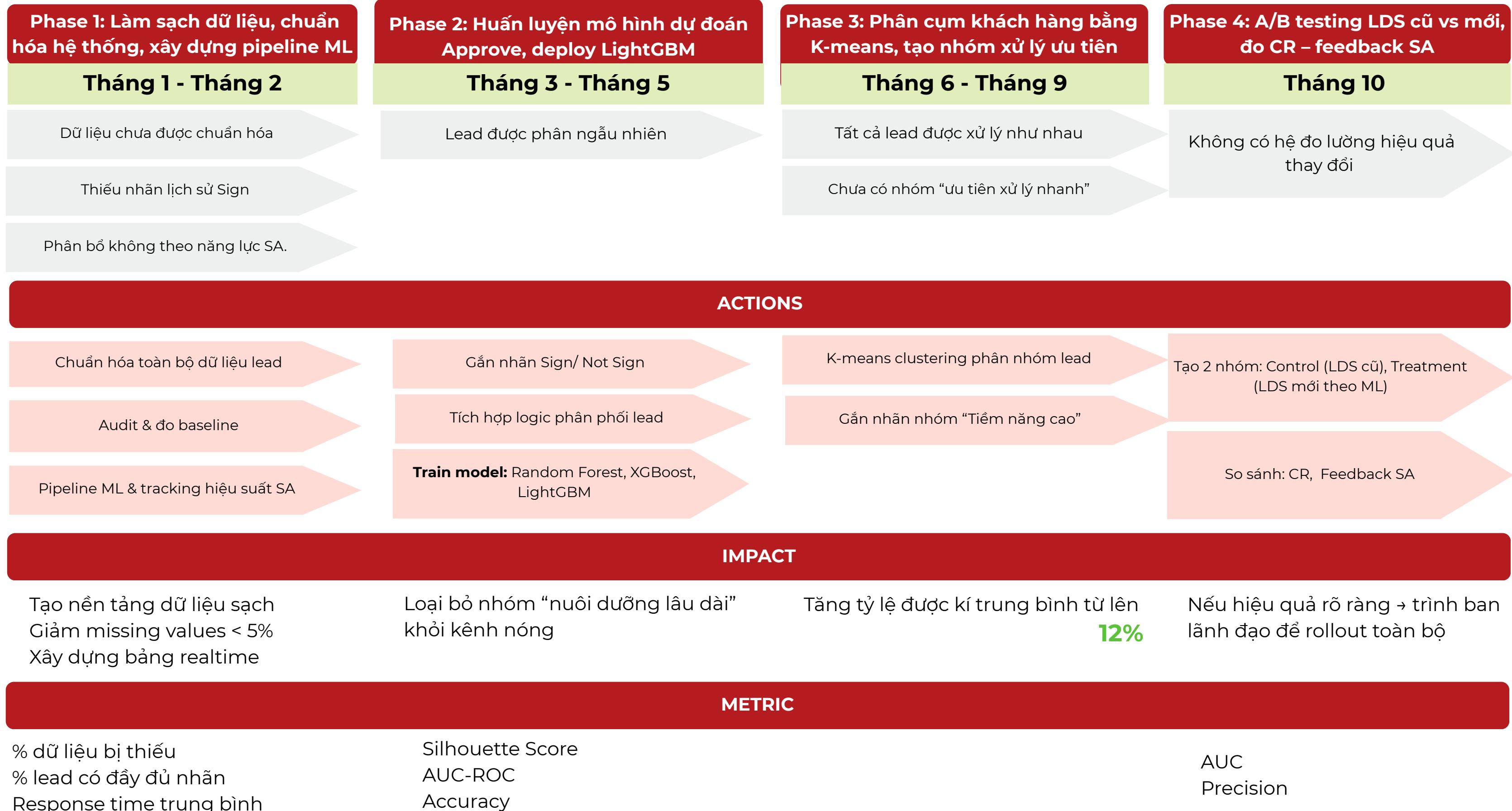


**Phase 5:** Rollout toàn hệ thống + tối ưu logic phân phối

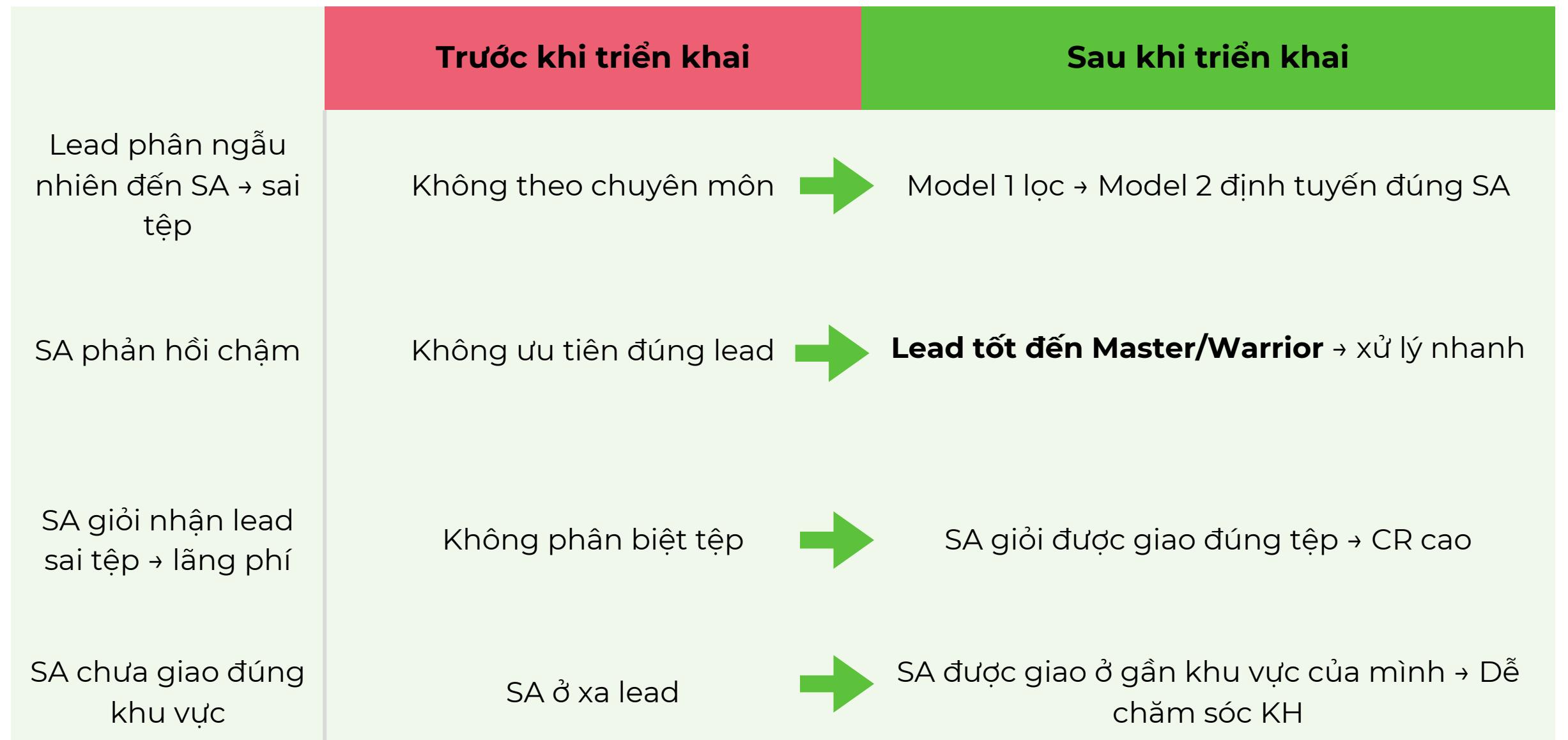


Hình Phân phối Leads theo tháng

# CỤ THỂ KẾ HOẠCH TRIỂN KHAI DỰ ÁN “SMART LEADS DISTRIBUTION SYSTEM”



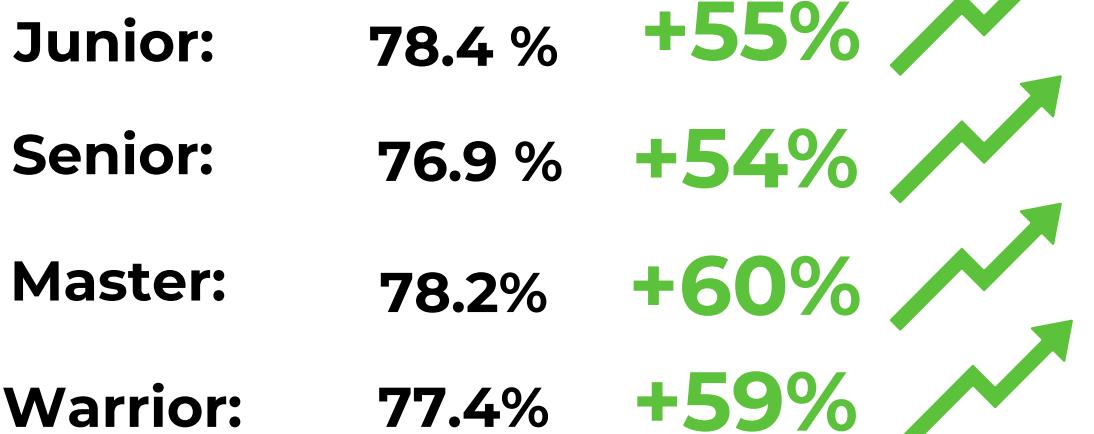
# SAU KHI TRIỂN KHAI DỰ ÁN “SMART LEADS DISTRIBUTION SYSTEM”



Conversion Rate: 68%

+33 %

Conversion Rate theo SA\_LEVEL

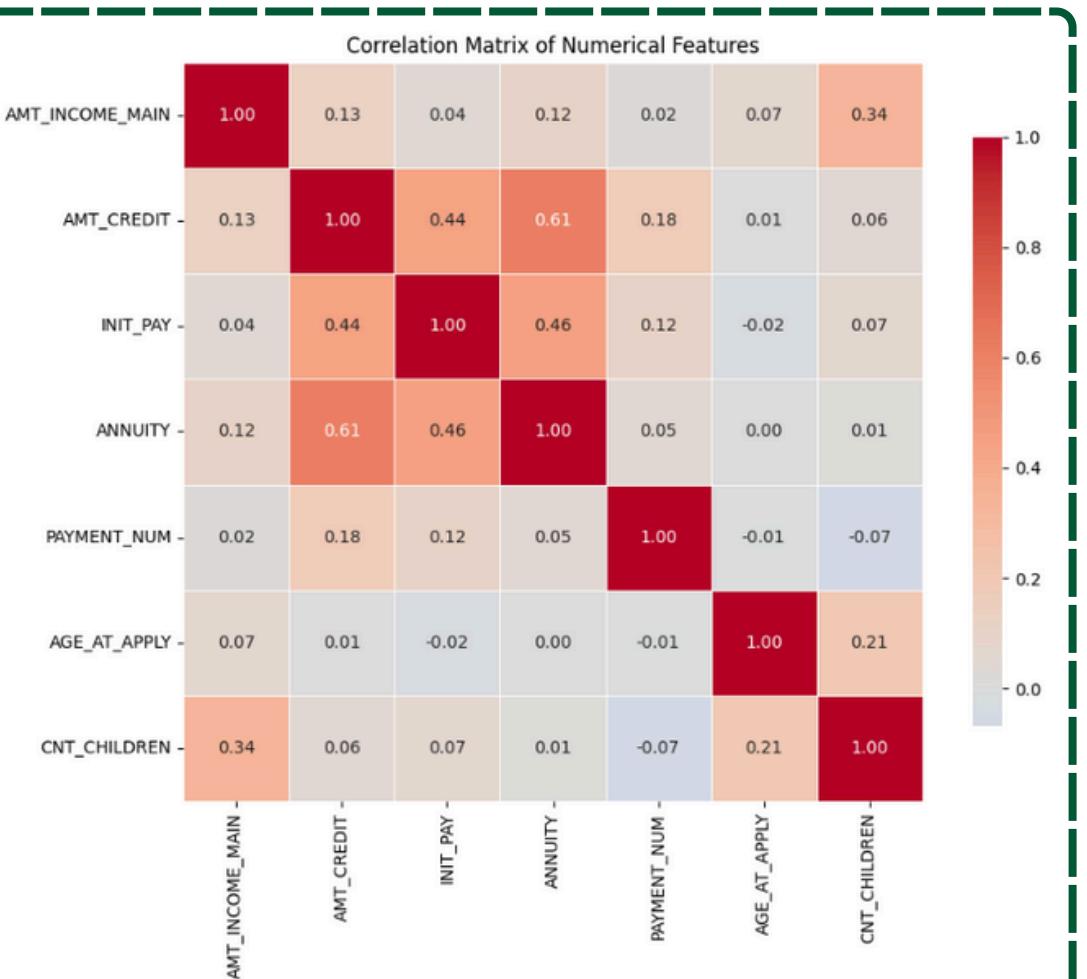


**CẢM ƠN MỌI NGƯỜI ĐÃ LẮNG NGHE**

**D2EQ**

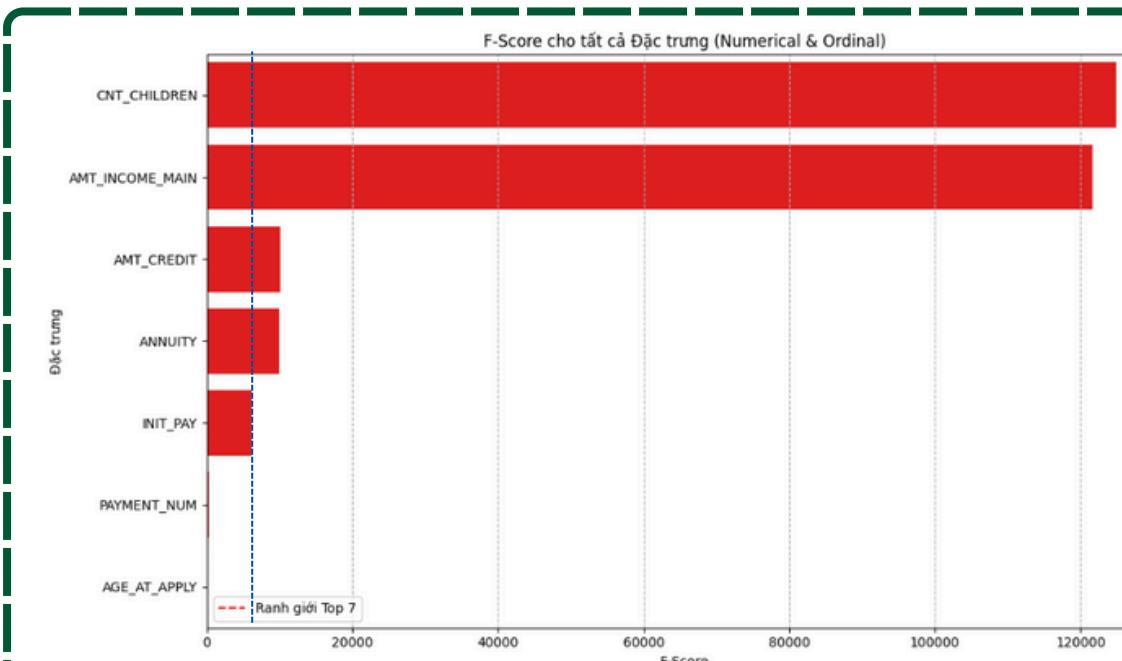
---

# Appendix Tối Ưu Hóa Tập Đặc Trưng: Xử lý Rò rỉ Dữ liệu & Lựa chọn Thống kê



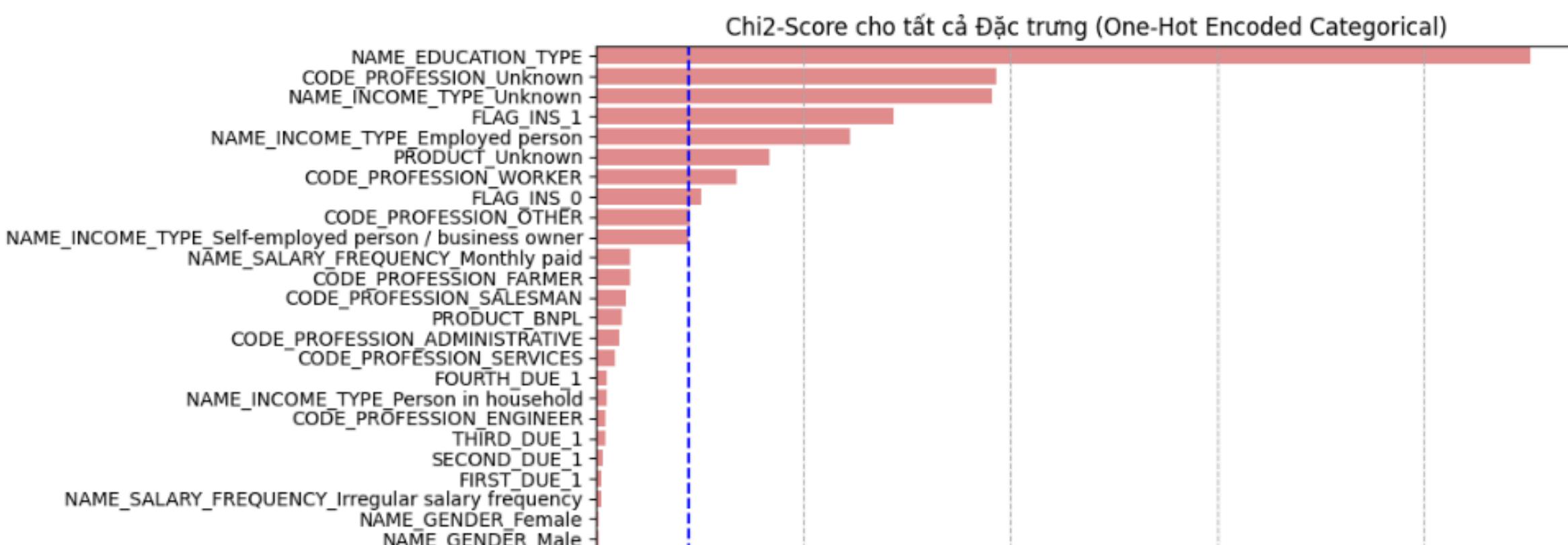
## Phân Tích Đặc Trưng Số:

- Báo cáo cho thấy không có cặp đặc trưng số nào có hệ số tương quan (correlation) vượt ngưỡng **0.8**. Điều này chỉ ra rằng không tồn tại hiện tượng **đa cộng tuyến** (multicollinearity) nghiêm trọng giữa các biến số trong tập dữ liệu. Do đó, giữ lại các đặc trưng số để tiếp tục chọn ra những đặc trưng tốt nhất



Đối với đặc trưng số:

- Đo lường mức độ khác biệt của đặc trưng giữa các nhóm khách hàng (được duyệt vs. không được duyệt).
- Chọn Top 5 đặc trưng có F-score cao nhất (ảnh hưởng mạnh nhất).



Đối với đặc trưng phân loại:

- Đo lường mối quan hệ phụ thuộc giữa đặc trưng phân loại và biến mục tiêu.
- Chọn Top 10 đặc trưng có Chi2-Score cao nhất (mối liên hệ mạnh nhất).

## Appendix

### Tinh chỉnh mô hình LightBGM

```
import lightgbm as lgb
from sklearn.metrics import classification_report, roc_auc_score, accuracy_score

# Tính toán scale_pos_weight
n_class0 = len(y_train[y_train == 0])
n_class1 = len(y_train[y_train == 1])
scale_pos_weight = n_class0 / n_class1

params = {
    'objective': 'binary',
    'metric': ['auc', 'binary_logloss'],
    'scale_pos_weight': scale_pos_weight,
    'num_leaves': 24,
    'min_child_samples': 300,
    'learning_rate': 0.03,
    'feature_fraction': 0.6,
    'bagging_fraction': 0.8,
    'bagging_freq': 5,
    'reg_alpha': 1.0,
    'reg_lambda': 1.0,
    'max_depth': 6,
    'min_split_gain': 0.1,
    'verbosity': -1,
    'seed': 42
}
```

# Appendix

```
► signed_statuses = ['Signed', 'Active', 'Finished', 'Paid off', 'Written off']

results_df['IS_SIGNED'] = results_df['NAME_CREDIT_STATUS'].isin(signed_statuses).astype(int)

signed_approved = results_df[(results_df['TARGET_PREDICTED'] == 1) & (results_df['IS_SIGNED'] == 1)].shape[0]
cr_model1 = signed_approved / len(results_df)
```

+ Code

+ Markdown

[89]:

cr\_model1

[89... 0.581332116161393

## Appendix

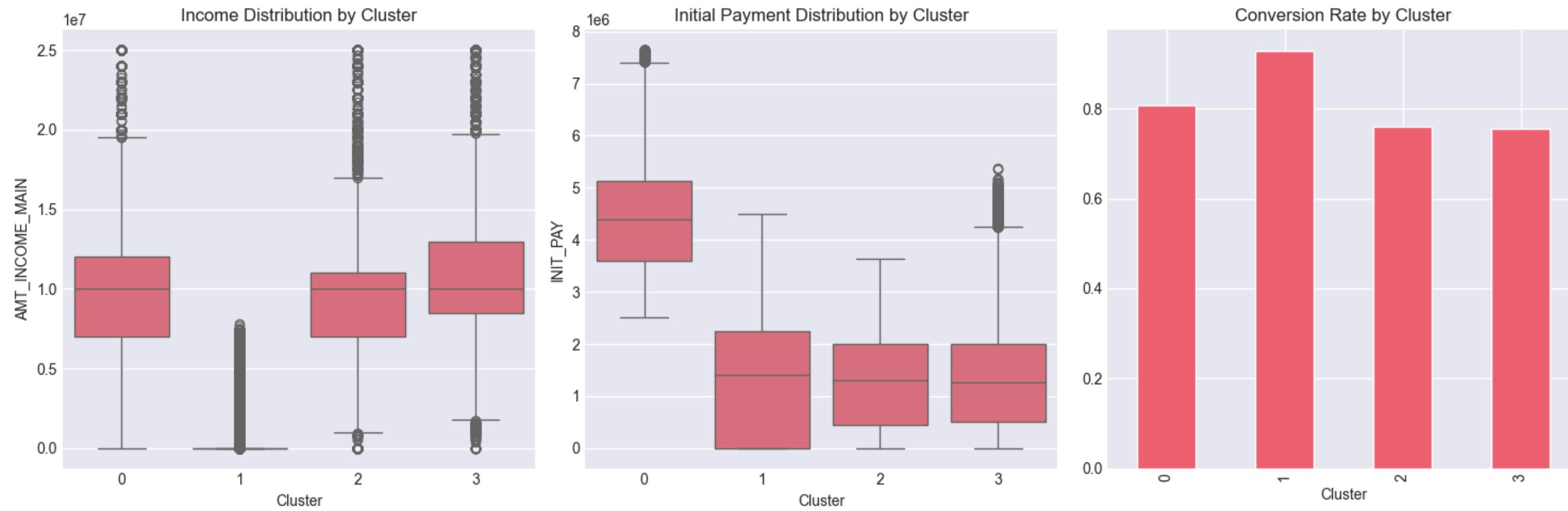
### Cluster Profiles:

CLUSTER	APPROVAL_PROBABILITY	AMT_CREDIT	AGE_AT_APPLY	AMT_INCOME_MAIN	\
0	0.51	8239246.26	34.75	11331532.23	
1	0.84	6457963.84	33.96	1.00	
2	0.76	2921441.40	34.23	7527394.13	
3	0.04	1.00	30.25	10536770.01	
CLUSTER	ANNUITY	NAME_INCOME_TYPE	CREDIT_INCOME_RATIO	DOWN_PAYMENT_RATIO	\
0	1409822.13	Employed person	25.49	0.33	
1	1138340.60	Unknown	6457963.84	0.40	
2	831060.28	Employed person	261794.30	4.90	
3	991000.00	Unknown	0.03	1611926.11	
CLUSTER	NAME_EDUCATION_TYPE	PRODUCT			
0	Unknown	Consumer Durable			
1	Unknown	Consumer Durable			
2	High school education	Consumer Durable			
3	Unknown	Consumer Durable			

# Appendix

Mục tiêu là gộp nhóm  
khách hàng để phân loại  
cho SA phù hợp

Có thể thấy thu nhập  
trung bình của cụm 2,3 là  
cao nhất trong các cụm.  
Tuy vậy tỉ lệ chuyển đổi lại  
khá thấp



Đưa cho SA có nhiều kinh nghiệm để tăng tỉ lệ chuyển đổi.

# Appendix

```
# Huấn luyện mô hình LightGBM cuối cùng với các siêu tham số tốt nhất
print("\n--- Huấn luyện mô hình LightGBM cuối cùng với các siêu tham số tốt nhất từ Optuna ---")
final_lgbm_model = lgb.LGBMClassifier(objective='binary', metric='auc', random_state=42, n_jobs=-1,
                                       **best_params)

final_lgbm_model.fit(X_train_selected_stat_top_N, y_train) # Huấn luyện trên toàn bộ tập huấn luyện

print("\nĐã huấn luyện mô hình LightGBM cuối cùng.")

# --- Đánh giá Mô hình LightGBM (Tính chính) trên Tập TEST ---
print("\n--- Đánh giá Mô hình LightGBM (Tính chính) trên Tập TEST (Sử dụng các chỉ số bạn muốn) ---")

# Dự đoán trên tập TEST
y_pred_test_final_lgbm = final_lgbm_model.predict(X_test_selected_stat_top_N)
y_proba_test_final_lgbm = final_lgbm_model.predict_proba(X_test_selected_stat_top_N)[:, 1]

# Tính toán và in ra các chỉ số mong muốn
auc_test_final_lgbm = roc_auc_score(y_test, y_proba_test_final_lgbm)
acc_test_final_lgbm = accuracy_score(y_test, y_pred_test_final_lgbm)

print(f"AUC-ROC (Test): {auc_test_final_lgbm:.4f}")
print(f"Độ chính xác (Accuracy) (Test): {acc_test_final_lgbm:.4f}")
print("Báo cáo phân loại (Test):")
print(classification_report(y_test, y_pred_test_final_lgbm))

# Tính toán F1-score cho lớp 0
report_test_final = classification_report(y_test, y_pred_test_final_lgbm, output_dict=True)
f1_class0_test_final = report_test_final['0']['f1-score']
print(f"F1-score lớp 0: {f1_class0_test_final:.4f}")

# Sử dụng StratifiedKFold để duy trì tỷ lệ lớp
kf = StratifiedKFold(n_splits=3, shuffle=True, random_state=42)
oof_preds = np.zeros(len(X_train_selected_stat_top_N))

list_test_auc = []
for fold, (train_index, val_index) in enumerate(kf.split(X_train_selected_stat_top_N, y_train)):
    X_train_fold, X_val_fold = X_train_selected_stat_top_N.iloc[train_index], X_train_selected_stat_top_N.iloc[val_index]
    y_train_fold, y_val_fold = y_train.iloc[train_index], y_train.iloc[val_index]

    model_lgbm = lgb.LGBMClassifier(**params) # Sử dụng LGBMClassifier của scikit-learn API
    model_lgbm.fit(X_train_fold, y_train_fold,
                   eval_set=[(X_val_fold, y_val_fold)],
                   eval_metric='auc',
                   callbacks=[lgb.early_stopping(stopping_rounds=50, verbose=False)])

    # Lấy AUC trên tập validation (tương ứng với test set của fold đó)
    y_val_proba = model_lgbm.predict_proba(X_val_fold)[:, 1]
    fold_auc = roc_auc_score(y_val_fold, y_val_proba)
    list_test_auc.append(fold_auc)

return np.mean(list_test_auc)
```