**QUESTION 6: MÔ HÌNH PHÂN KHÚC VÀ DỰ ĐOÁN RỜI BỎ KHÁCH HÀNG CHO HIGHLANDS COFFEE**

1. **Tổng quan và mục tiêu**

Báo cáo này trình bày một quy trình phân tích dữ liệu kết hợp mô hình phân khúc khách hàng và dự đoán rời bỏ (churn prediction), từ đó giúp Highlands Coffee thiết lập các chiến lược giữ chân khách hàng hiệu quả hơn.

Giải pháp được phát triển từ dữ liệu khách hàng thực tế trên nhiều chiều: hành vi, cảm nhận thương hiệu, nhân khẩu học, ngữ cảnh sử dụng, v.v., với mục tiêu cuối cùng là:

* Nhận diện các phân khúc khách hàng chính
* Dự đoán nguy cơ rời bỏ theo từng khách
* Đề xuất hướng hành động giữ chân phù hợp theo nhóm

1. **Xử lý dữ liệu và tạo đặc trưng(Feature Engineering)**

Đây là bước đầu tiên và quan trọng trong bất kỳ pipeline phân tích nào. Mục tiêu là tạo một bảng dữ liệu đầu vào hoàn chỉnh, đầy đủ đặc trưng phản ánh hành vi, cảm nhận, nhu cầu và nhân khẩu học của khách hàng để phục vụ cho các bước mô hình hóa phía sau.

* 1. **Tổng hợp dữ liệu từ nhiều bảng**

DataFrame ban đầu được merge từ nhiều bảng: SA\_var\_cleaned, Brand\_Health, Segmentation, Brand\_Image, Companion, Dayofweek, Daypart, Needstate. Các biến được đồng bộ, chọn đặc trưng numeric hoặc lấy giá trị đầu tiên theo ID.

Dữ liệu sau merge có dạng mỗi ID một dòng duy nhất

* 1. **Các nhóm đặc trưng được tạo**

*Perceptual*: comprehension, brand\_likability, nps\_p3m, nps\_p3m\_group

*Behavioral*: visit, ppa, spending, visit#dayofweek, visit#daypart

*Segmentation*: segmentation

*Companionship*: companion#group

*Need states*: needstategroup, needstates

*Demographics*: Age, Gender, Occupation, MPI#2

* 1. **Xử lý dữ liệu thiếu và encode**

Dữ liệu số: điền median

Dữ liệu perception thiếu → gán 0 (không biết/đánh giá)

Các biến object được LabelEncoded

**A screen shot of a computer

AI-generated content may be incorrect.**

* 1. **Tạo biến churn**

Churn = 1 nếu khách hàng có hoạt động trong 3 tháng gần nhất (p3m = 1) NHƯNG KHÔNG đến trong tháng gần nhất (p1m = 0). Ngược lại là Retained.

1. **Mô hình phân khúc khách hàng**

Phân khúc khách hàng giúp doanh nghiệp hiểu rõ sự đa dạng trong tập khách hàng, từ đó đưa ra chiến lược cá nhân hóa phù hợp theo từng nhóm. Mô hình KMeans được chọn vì đơn giản, trực quan và phù hợp với dữ liệu chuẩn hóa tuyến tính.

* 1. **Tiêu chí chọn biến phân cụm**

Chọn các biến không phụ thuộc churn: comprehension, brand\_likability, visit, ppa, Age, MPI#2, segmentation, needstategroup

* 1. **Phương pháp phân cụm**
* Biểu đồ Elbow Method biểu diễn mối quan hệ giữa số lượng cụm (k) và inertia (tổng khoảng cách nội cụm). Điểm gấp rõ nhất tại k=4 cho thấy đây là số cụm tối ưu vì sau đó độ cải thiện giảm dần.

A graph with a line

AI-generated content may be incorrect.

* Dù k=2 có điểm silhouette cao nhất, nhưng segmentation như vậy quá ít phân hóa. K=4 là mức dung hòa giữa phân hóa đủ sâu và độ ổn định tương đối (score ~0.318).
  + Standardize: StandardScaler
  + Dùng **Elbow method** và **Silhouette Score** để chọn k = 4
  + Mô hình: KMeans(n\_clusters=4)
  + Gán nhãn segment\_cluster cho từng khách hàng

A black background with white text

AI-generated content may be incorrect.

* 1. **Huấn luyện mô hình Kmeans và gán cluster**

Chuẩn hóa dữ liệu

Dùng KMeans(n\_clusters=4)

Gán nhãn cluster vào X, df

* 1. **Phân tích profile cluster**
* Trung bình hành vi, perception, MPI, Age...
* Nhận diện nhóm trung thành, rủi ro, tiềm năng cao

A screenshot of a graph

AI-generated content may be incorrect.

Biểu đồ này cho thấy trung bình giá trị các đặc trưng theo từng cụm khách hàng. Ví dụ: cụm số 3 có chi tiêu (spending) rất cao và số lần đến quán lớn (visit ~32), gợi ý đây là nhóm khách hàng trung thành/cao cấp. Cụm 0 và 1 có hành vi và perception thấp, có thể là nhóm rủi ro hoặc ít quan tâm đến thương hiệu.

1. **Mô hình dự đoán rời bỏ(Churn Prediction)**

Mục tiêu chính là xác định ai có khả năng rời bỏ trong tương lai gần để có hành động giữ chân sớm. Mô hình Random Forest phù hợp do khả năng xử lý nhiều loại biến và cho phép phân tích độ quan trọng của từng đặc trưng.

* 1. **Định nghĩa mục tiêu**

 churn là biến nhị phân: 0 hoặc 1

 Mục tiêu là dự đoán khả năng churn dựa vào các đặc trưng

* 1. **Chọn đặc trưng đầu vào**

Bao gồm:

* Cảm nhận thương hiệu
* Hành vi
* Phân khúc
* Kết quả cụm (cluster)
* Nhu cầu
  1. **Xử lý mất cân bằng**

Churn = 1 chiếm tỉ lệ thấp → dùng upsampling để cân bằng tập train

* 1. **Huấn luyện mô hình**

 Sử dụng **Random Forest**

 **GridSearchCV** để tối ưu hyperparameter

|  |
| --- |
| Best Params: {'max\_depth': 10, 'min\_samples\_split': 2, 'n\_estimators': 100} |

 Chia train/test 80/20

* 1. **Kết quả và đánh giá**
* Accuracy: **100%**
* F1-score: **100%**
* Precision/Recall (class 1): **1.00 / 1.00**

**A screenshot of a computer program

AI-generated content may be incorrect.**

Mô hình đạt kết quả hoàn hảo trên tập test (100% chính xác), điều này có thể là tín hiệu tốt nhưng cũng cần kiểm tra kỹ hơn về overfitting và độ đa dạng của dữ liệu.

* 1. **Phân tích các yếu tố ảnh hưởng**

Việc hiểu rõ lý do khiến khách hàng rời bỏ là nền tảng cho việc xây dựng chiến lược giữ chân hiệu quả. Phân tích feature importance từ Random Forest giúp xác định đâu là yếu tố tác động nhiều nhất đến churn.

* Trích xuất feature\_importance\_ từ Random Forest
* Các yếu tố quan trọng nhất:
  + Điểm brand\_likability thấp
  + Thuộc cluster rủi ro
  + Hành vi đến ít
  + Đi một mình (không đi nhóm)

A graph with blue and white bars

AI-generated content may be incorrect.

Biểu đồ cho thấy các biến có ảnh hưởng lớn nhất đến dự đoán rời bỏ khách hàng. Top 3 bao gồm ppa\_y (mức chi tiêu trung bình mỗi lần đến), spending\_y (tổng chi tiêu), và Occupation (nghề nghiệp). Điều này cho thấy nhóm khách có hành vi chi tiêu cao, hoặc đặc trưng nghề nghiệp nhất định, có xác suất rời bỏ khác biệt rõ. Một số yếu tố perception như brand\_likability hay comprehension lại có tầm ảnh hưởng rất thấp trong mô hình này.

1. **Hạn chế và hướng phát triểu tương lai**
   1. **Hạn chế**

 Một số khách hàng thiếu dữ liệu perception (comprehension, nps) → ảnh hưởng phân cụm.

 Việc gán giá trị 0 cho perception có thể gây bias (vì 0 mang hàm ý “không thích” chứ không phải “không có dữ liệu”).

 Dữ liệu chỉ trong một thời điểm → chưa bắt được thay đổi hành vi theo thời gian.

* 1. **Đề xuất tương lai**

 Thu thập thêm dữ liệu hành vi theo thời gian thực (longitudinal).

 Sử dụng thêm thuật toán phân cụm nâng cao như DBSCAN hoặc Gaussian Mixture Model.

 Thử nghiệm các thuật toán churn prediction khác như XGBoost, LightGBM.

 Phân tích bổ sung churn theo từng cụm (segment\_cluster) để xây dựng retention plan riêng biệt.