E-commerce

Classification

Customer Subscription Prediction

Xây dựng mô hình dự đoán xác suất đăng ký dịch vụ của khách hàng

Nguyễn Diệu Linh
K6 - ML

NỘI DUNG

01. Giới thiệu & mục tiêu bài toán

- **Thách thức hiện tại:** Cạnh tranh gay gắt, khách hàng có nhiều lựa chọn, tỷ lệ đăng ký dịch vụ giá trị gia tăng còn khiêm tốn.

- **Giải pháp:** Xây dựng mô hình machine learning dự đoán khả năng khách hàng đăng ký dịch vụ.

- **Lợi ích:**
  - Tối ưu hóa chiến lược marketing, tập trung vào nhóm khách hàng có xác suất đăng ký cao.
  - Tăng trưởng doanh thu bền vững từ khách hàng trung thành, mua lặp lại, đóng góp doanh thu định kỳ.
  - Nâng cao trải nghiệm khách hàng qua cá nhân hóa, xây dựng mối quan hệ lâu dài.

02. Tổng quan dữ liệu & Insights

Tổng quan dữ liệu

Nguồn dữ liệu: Consumer Behavior and Shopping Habits Dataset trên Kaggle.

Tập dữ liệu gốc: `shopping_behavior_updated.csv` (3900 dòng x 18 cột).

- Thông tin về nhân khẩu học, lịch sử mua hàng, sở thích và kênh mua sắm của khách hàng.
- **Target**: Subscription Status (tình trạng đăng ký).

Chất lượng dữ liệu: **không** có dữ liệu bị thiếu (null) hay lặp lại (duplicate).

Variable Name	Role	Type	Description
Customer_ID	Feature	Integer	Unique identifier for each customer, enabling tracking and analysis of their shopping behavior over time.
Age	Feature	Integer	The customer's age, providing demographic information for segmentation and targeted marketing.
Gender	Feature	Categorical	The customer's gender, a key demographic factor influencing product preferences and purchasing patterns.
Item_Purchased	Feature	Categorical	The specific item chosen by the customer during the transaction.
Category	Feature	Categorical	The broad group to which the purchased item belongs (e.g., clothing, electronics, groceries).
Purchase_Amount_USD	Feature	Integer	The transaction's monetary value in U.S. dollars, indicating the cost of the purchased item(s).
Location	Feature	Categorical	The geographical region where the purchase was made, offering insights into regional preferences and market trends.
Size	Feature	Categorical	The size specification (if applicable) of the purchased item, relevant for apparel, footwear, and certain consumer goods.
Color	Feature	Categorical	The color variant of the purchased item, influencing customer preferences and stock availability.
Season	Feature	Categorical	Seasonal relevance of the purchased item (spring, summer, fall, winter), impacting inventory management and marketing strategies.
Review_Rating	Feature	Float	Numerical assessment provided by the customer regarding their satisfaction with the purchased item.
Shipping_Type	Feature	Categorical	Delivery method used (e.g., standard shipping, express delivery), influencing delivery time and cost.
Discount_Applied	Feature	Binary	Indicates whether a promotional discount was used, shedding light on price sensitivity and promotion effectiveness.
Promo_Code_Used	Feature	Binary	Indicates whether a coupon or promo code was applied, aiding evaluation of marketing campaign success.
Previous_Purchases	Feature	Integer	Number of prior purchases made by the customer, contributing to customer segmentation and retention strategies.
Payment_Method	Feature	Categorical	Mode of payment used by the customer (e.g., credit card, cash), offering insights into preferred payment options.
Frequency_of_Purchases	Feature	Categorical	How often the customer makes purchases (e.g., daily, weekly, monthly), a critical metric for assessing loyalty and lifetime value.
Subscription_Status	**Target**	**Binary**	Indicates whether the customer has opted for a subscription service, offering insights into loyalty and potential for recurring revenue (Yes/No).

Source: *Customer Behavior and Shopping Habits Dataset.* (n.d.). Kaggle.com. Retrieved June 2025, from https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset

02. Tổng quan dữ liệu & Insights

Tổng quan dữ liệu

Số liệu thống kê cơ bản của các biến dạng số (numerical features):

	Customer_ID	Age	Purchase_Amount_(USD)	Review_Rating	Previous_Purchases
count	3900.000000	3900.000000	3900.000000	3900.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.749949	25.351538
std	1125.977353	15.207589	23.685392	0.716223	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.700000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

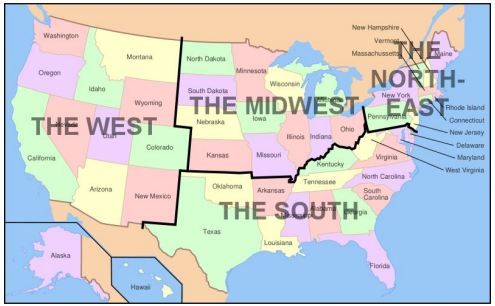
(Loại bỏ)

02. Tổng quan dữ liệu & Insights

Tổng quan dữ liệu

Số lượng giá trị unique của các biến phân loại (categorical features):

	Name	Number_Of_Unique_Value	Unique_Value_Name
0	Gender	2	Male, Female
1	Item_Purchased	25	Blouse, Sweater, Jeans, Sandals, Sneakers, Shirt, Shorts, Coat, Handbag, Shoes, Dress, Skirt, Sunglasses, Pants, Jacket, Hoodie, Jewelry, T-shirt, Scarf, Hat, Socks, Backpack, Belt, Boots, Gloves
2	Category	4	Clothing, Footwear, Outerwear, Accessories
3	Location	50	Kentucky, Maine, Massachusetts, Rhode Island, Oregon, Wyoming, Montana, Louisiana, West Virginia, Missouri, Arkansas, Hawaii, Delaware, New Hampshire, New York, Alabama, Mississippi, North Carolina, California, Oklahoma, Florida, Texas, Nevada, Kansas, Colorado, North Dakota, Illinois, Indiana, Arizona, Alaska, Tennessee, Ohio, New Jersey, Maryland, Vermont, New Mexico, South Carolina, Idaho, Pennsylvania, Connecticut, Utah, Virginia, Georgia, Nebraska, Iowa, South Dakota, Minnesota, Washington, Wisconsin, Michigan
4	Size	4	L, S, M, XL
5	Color	25	Gray, Maroon, Turquoise, White, Charcoal, Silver, Pink, Purple, Olive, Gold, Violet, Teal, Lavender, Black, Green, Peach, Red, Cyan, Brown, Beige, Orange, Indigo, Yellow, Magenta, Blue
6	Season	4	Winter, Spring, Summer, Fall
7	Shipping_Type	6	Express, Free Shipping, Next Day Air, Standard, 2-Day Shipping, Store Pickup
8	Discount_Applied	2	Yes, No
9	Promo_Code_Used	2	Yes, No
10	Payment_Method	6	Venmo, Cash, Credit Card, PayPal, Bank Transfer, Debit Card
11	Frequency_of_Purchases	7	Fortnightly, Weekly, Annually, Quarterly, Bi-Weekly, Monthly, Every 3 Months

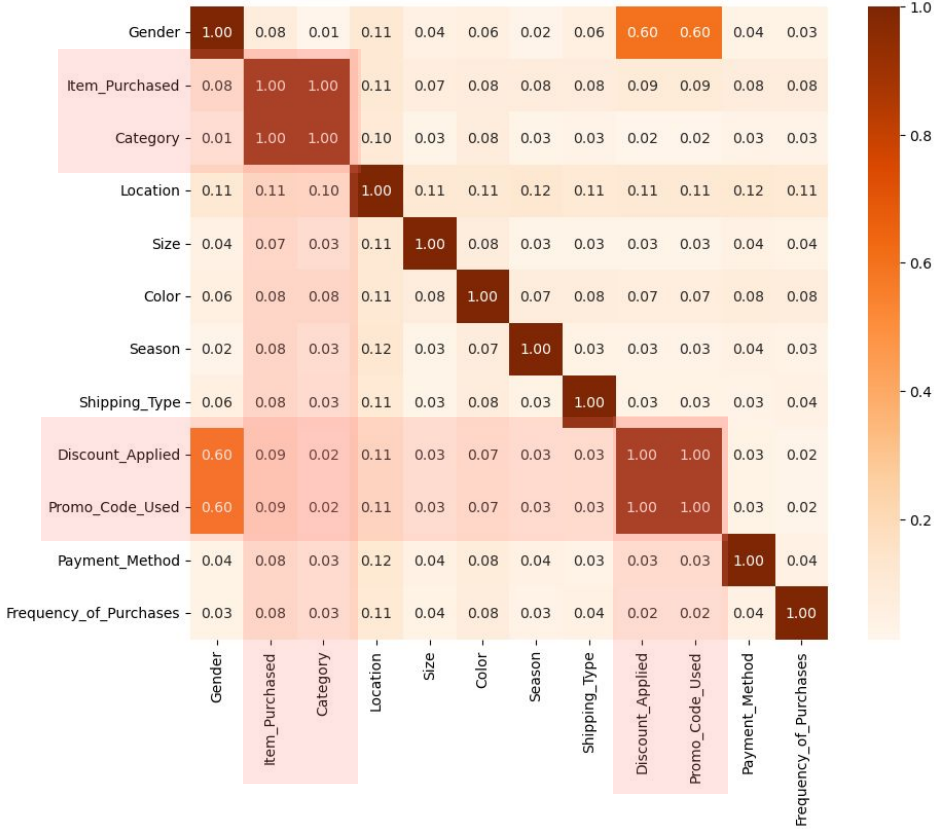


Nhóm thành 4 khu vực
(vì đây là các tiểu bang của Mỹ)

Nhóm thành 3 nhóm màu:
cool, warm và neutral

02. Tổng quan dữ liệu & Insights

Tổng quan dữ liệu



Item Purchased có tương quan hoàn hảo với **Category** (do mỗi sản phẩm đều thuộc một nhóm category cụ thể), nên loại bỏ **Item Purchased** để tránh trùng lặp thông tin và làm mô hình gọn hơn.

Tương tự với trường hợp của 2 biến **Discount Applied** và **Promo Code Used**, chúng có cùng ý nghĩa là sản phẩm hay đơn hàng có được giảm giá hay không. Vì vậy, chỉ giữ lại biến **Promo Code Used.**

(Loại bỏ Item Purchased và Discount Applied)

02. Tổng quan dữ liệu & Insights

Tổng quan dữ liệu

Tập dữ liệu sử dụng: 3900 dòng x 15 cột, trong đó có 4 biến dạng số (numerical features), 10 biến dạng phân loại (categorical features) và 1 biến mục tiêu (target).

Numerical features	Age, Purchase_Amount_(USD), Review_Rating, Previous_Purchases
Categorical features	Gender, Category, Size, Season, Shipping_Type, Promo_Code_Used, Payment_Method, Frequency_of_Purchases, Region, Color_Group
Target	Subscription_Status

Dữ liệu về target bị mất cân bằng, với 73% dữ liệu nằm ở lớp 0 (không đăng ký) và chỉ có 27% dữ liệu nằm ở lớp 1 (đăng ký).

02. Tổng quan dữ liệu & Insights

Insights

Thông tin về đăng ký dịch vụ:

- **Tỷ lệ đăng ký dịch vụ còn thấp**: Có tới 73% khách hàng không đăng ký dịch vụ.
- **Khác biệt rõ rệt theo giới tính:** Toàn bộ khách hàng đăng ký đều là nam.
- **Ảnh hưởng của chương trình giảm giá:** Việc đăng ký dịch vụ chỉ diễn ra khi có chương trình giảm giá.

Đặc điểm:

- **Nhóm tuổi chủ đạo**: Khách hàng chủ yếu là người trung niên và cao tuổi, đây cũng là nhóm chi tiêu nhiều và có xu hướng mua lặp lại thường xuyên hơn.
- **Giá trị đơn hàng:** Hầu hết các giao dịch tập trung vào hai khoảng giá: trung bình (20–40 USD) và cao (90–100 USD).
- **Lựa chọn giao hàng:** Phần lớn khách hàng ưu tiên lựa chọn miễn phí vận chuyển.

02. Tổng quan dữ liệu & Insights

Insights

Sự khác biệt giữa khách hàng nam giới và nữ giới:

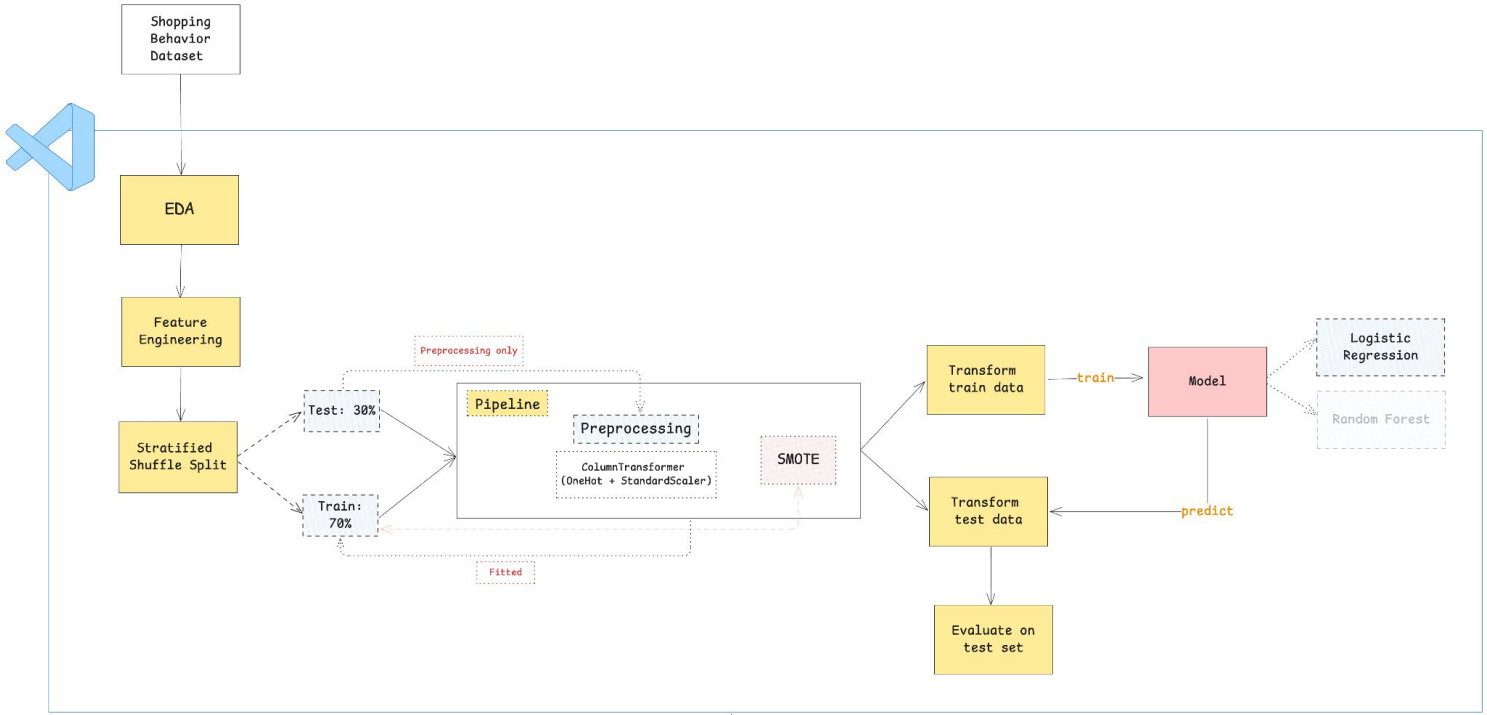
- **Cơ cấu khách hàng**: 68% khách hàng là nam giới.
- **Tần suất mua sắm:** Nam giới có thói quen mua sắm trung bình mỗi ba tháng một lần, trong khi nữ giới mua hàng thường xuyên hơn, khoảng hai lần mỗi tuần.
- **Ảnh hưởng của giảm giá:** Nam giới chỉ mua hàng khi có chương trình giảm giá, trong khi nữ giới vẫn mua ngay cả khi không có khuyến mãi.
- **Giá trị chi tiêu:** Tuy nữ giới chi tiêu trung bình mỗi đơn hàng cao hơn, nhưng tổng số tiền mua sắm của nam giới lại gần gấp đôi do số lượng khách hàng nam đông hơn và tần suất mua lặp lại cũng cao hơn.

Mức độ hài lòng của khách hàng:

- Điểm đánh giá (Review Rating) cho thấy mức độ hài lòng của khách hàng rất cao và tất cả khách hàng đều có ít nhất một lần mua lại.

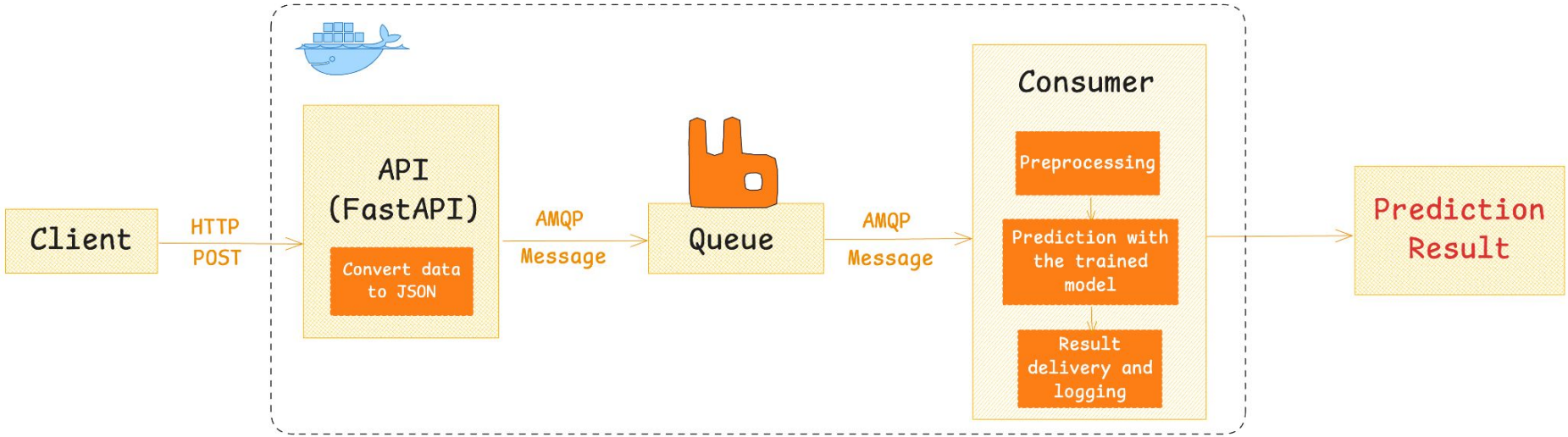
03. Phương pháp luận

Giai đoạn 1: Xây dựng mô hình



03. Phương pháp luận

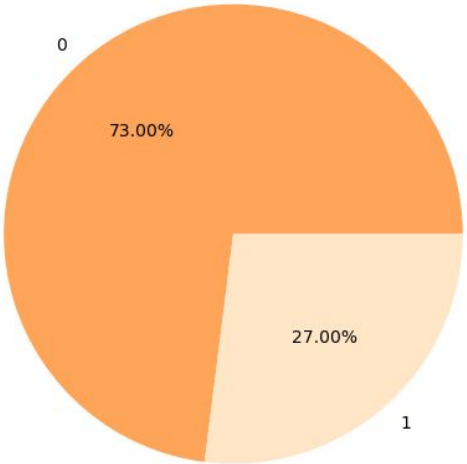
Giai đoạn 2: Triển khai mô hình



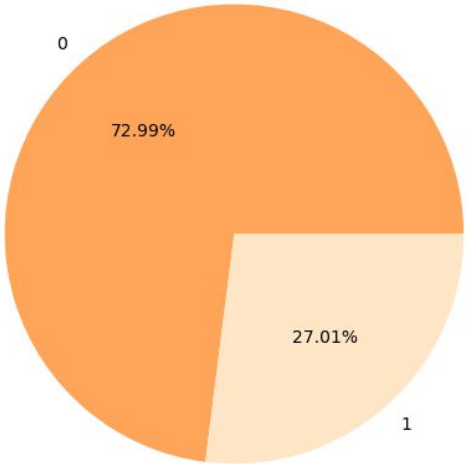
04. Thực hiện & Kết quả mô hình

Chia dữ liệu

Stratified Shuffle Split



Train dataset

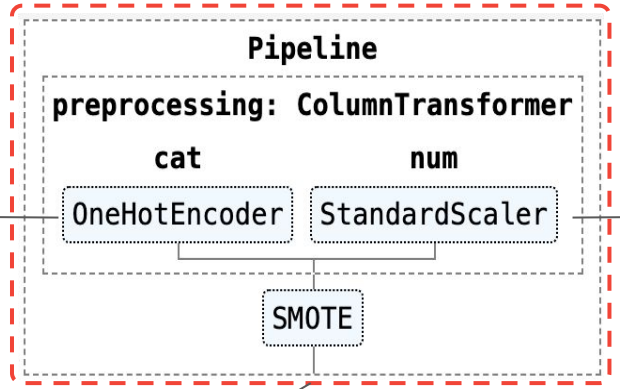


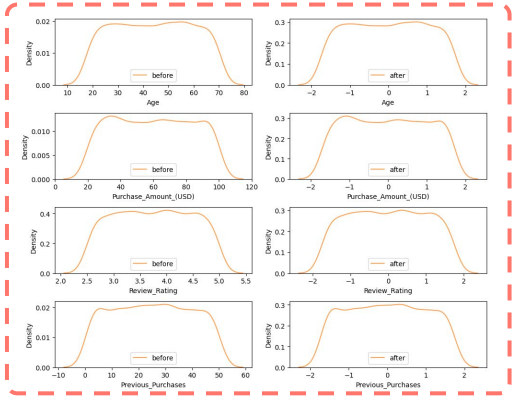
Test dataset

04. Thực hiện & Kết quả mô hình

Xây dựng Data Pipeline

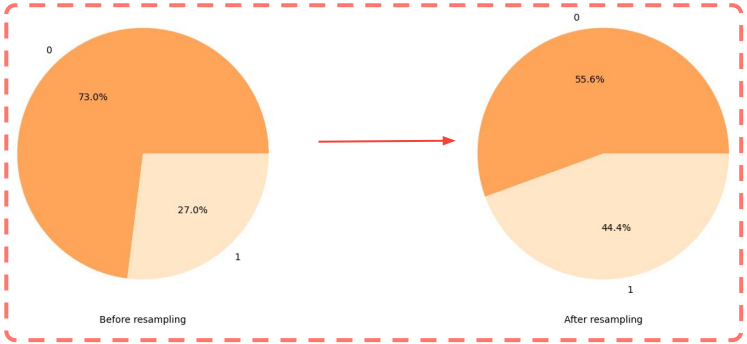
42 cột mới được sinh ra. Data sparsity ≈ 76%.





Distribution của các numerical features không thay đổi đáng kể

Cân bằng lại tỷ lệ giữa hai lớp trong tập train



04. Thực hiện & Kết quả mô hình

Xây dựng mô hình

Logistic Regression

Random Forest

ROC AUC	Hệ số Gini	Accuracy	Precision	Recall	F1-score

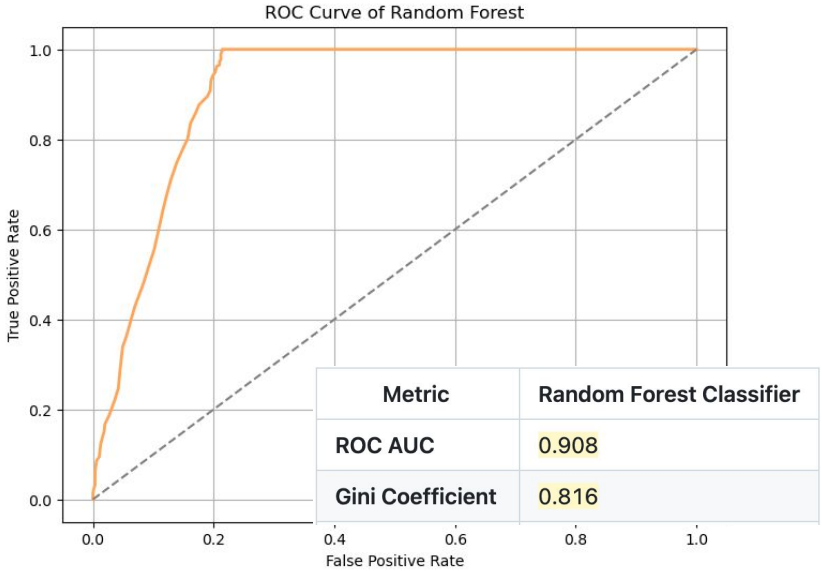
04. Thực hiện & Kết quả mô hình

Xây dựng mô hình

Logistic Regression

Metric	Logistic Regression
ROC AUC	0.893
Gini Coefficient	0.786

Random Forest



Metric	Random Forest Classifier
ROC AUC	0.908
Gini Coefficient	0.816

04. Thực hiện & Kết quả mô hình

Xây dựng mô hình

Model	AUC (train)	AUC (test)	Overfitting	Underfitting
Logistic Regression	0.913	0.893	Not overfitting	Not underfitting
Random Forest	1.000	0.908	Overfitting	Not underfitting

	PSI Score
Logistic Regression	0.0082 (LOW)
Random Forest	2.4896 (HIGH)

Từ những phân tích trên, lựa chọn Logistic Regression

04. Thực hiện & Kết quả mô hình

Xây dựng mô hình

Logistic Regression

Top 3 biến có giá trị coefficient dương cao nhất

Feature	Coefficient
cat__Promo_Code_Used_Yes	3.456289
cat__Gender_Male	0.765311
cat__Shipping_Type_Express	0.345605

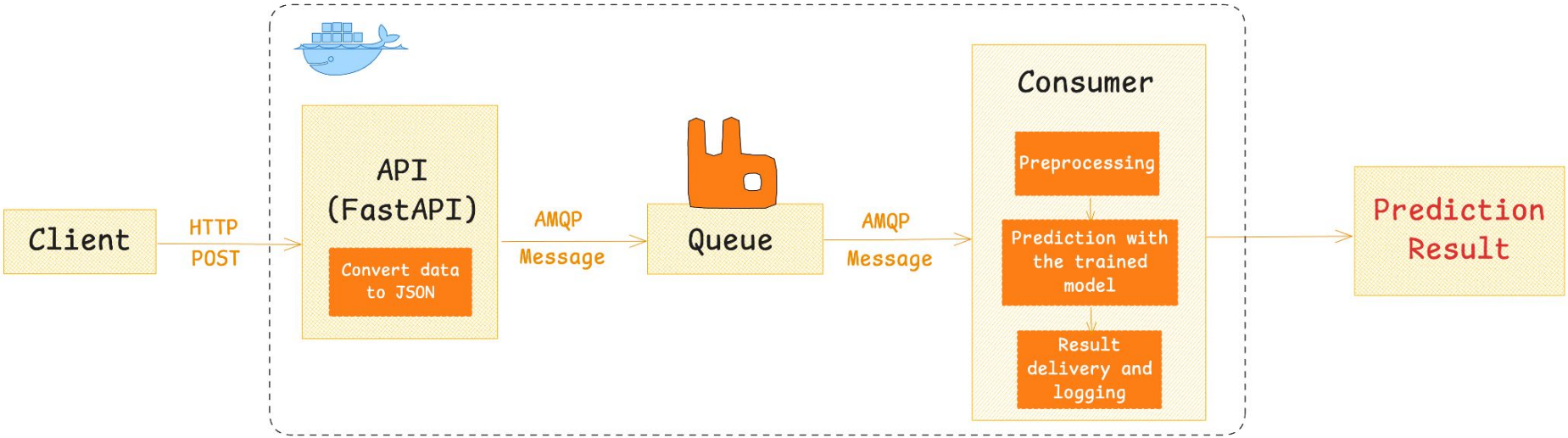
Các biến này tăng sẽ làm tăng xác suất xảy ra lớp 1

Top 3 biến có giá trị coefficient âm thấp nhất

Feature	Coefficient
cat__Promo_Code_Used_No	-3.457021
cat__Gender_Female	-0.766043
cat__Shipping_Type_Next Day Air	-0.362490

Các biến này tăng sẽ làm tăng xác suất xảy ra lớp 0

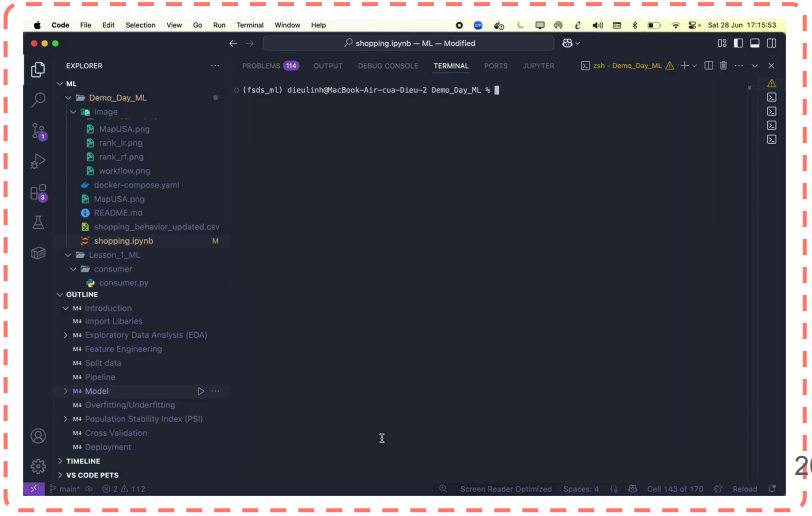
05. Triển khai mô hình



05. Triển khai mô hình

```
curl -X POST "http://localhost:8081/predict" \
  -H "Content-Type: application/json" \
  -d '{
    "Age": 35,
    "Gender": "Male",
    "Category": "Clothing",
    "Purchase_Amount_(USD)": 49,
    "Size": "M",
    "Season": "Spring",
    "Review_Rating": 3.7,
    "Shipping_Type": "Express",
    "Promo_Code_Used": "Yes",
    "Previous_Purchases": 5,
    "Payment_Method": "Venmo",
    "Frequency_of_Purchases": "Weekly",
    "Region": "Midwest",
    "Color_Group": "Warm"
  }'
```

```
🚀 Consumer run.
✅ Successfully connected to RabbitMQ!
Waiting for messages 🔍
Message received from queue: b'{"Age": 35, "Genc
✅ Probability: 0.80
✅ Label: Subcribed
```



06. Phần mở rộng

Đề xuất giải pháp cho bài toán

- Triển khai chương trình ưu đãi cho lần đầu đăng ký như mã giảm giá áp dụng ngay sau đăng ký.
- Duy trì ưu đãi/giảm giá vào dịp đặc biệt như sinh nhật, ngày lễ.
- Thường xuyên gửi mã giảm giá phí vận chuyển riêng cho những khách hàng đã đăng ký.
- Quảng bá các ưu đãi độc quyền cho các khách hàng đã đăng ký để tạo hiệu ứng FOMO.
- Doanh nghiệp nên tối ưu quy trình đăng ký đơn giản, dễ thao tác, giảm thiểu rào cản.

06. Phần mở rộng

Một số bài toán khác từ tập dataset

	Mục tiêu	Target	Thuật toán
Dự đoán giá trị đơn hàng (Purchase Value Prediction)	Dự đoán số tiền mà mỗi khách hàng sẽ tiêu trong các giao dịch tiếp theo.	Purchase Amount (USD)	Hồi quy
Phân nhóm khách hàng (Customer Clustering)	Phân nhóm khách hàng dựa trên hành vi nhân khẩu học, sở thích, đặc điểm	–	Unsupervised learning (vì chưa biết nhóm khách hàng)

CẢM ƠN ANH QUAN VÀ FSDS