TRƯỜNG PHÂN HIỆU ĐẠI HỌC THUỶ LỢI

CHUYÊN NGÀNH CÔNG NGHỆ THÔNG TIN

BÁO CÁO HỌC PHẦN: KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: 3A Superstore (Market Orders Data-CRM)

Giảng viên hướng dẫn:

TS. Vũ Thị Hạnh

Nhóm sinh viên thực hiện:

Hồ Thanh Phúc – 2251068233

Lê Thừa Lạc – 2251068203

TP Hồ Chí Minh, ngày 20 tháng 10 năm 2025

Họ và tên	Công việc thực hiện
Lê Thừa Lạc	- Phân tích doanh thu theo tháng, danh mục sản phẩm, khu vực. - Phân tích doanh thu danh mục Home theo từng khu vực. - Phân tích doanh thu sản phẩm Musical Heart Plush Bear 37 Cm từng khu vực. - Dự đoán danh thu sản phẩm Musical Heart Plush Bear 37 Cm trong 6 tháng tới - Dự đoán danh thu danh mục Home theo khu vực trong 6 tháng tới.
Hồ Thanh Phúc	- Dự đoán xác xuất khách sẽ mua hàng trong kỳ tiếp theo. - Dự đoán giá trị vòng đời của khách hàng. - Dự đoán khả năng rời bỏ của khách hang.

Mục Lục

MỤC LỤC HÌNH ẢNH

LỜI MỞ ĐẦU

Trong bối cảnh thị trường bán lẻ ngày càng cạnh tranh khốc liệt, việc thấu hiểu khách hàng và tối ưu hóa hoạt động kinh doanh dựa trên dữ liệu đang trở thành yếu tố sống còn đối với các doanh nghiệp. Mỗi giao dịch, mỗi sản phẩm được bán ra hay thậm chí mỗi hành vi mua sắm đều để lại "dấu vết dữ liệu" quý giá, giúp doanh nghiệp nhìn nhận sâu hơn về xu hướng tiêu dùng và hành vi khách hàng.

Dự án "3A Superstore (Market Orders Data – CRM)" được thực hiện với mục tiêu mô phỏng và phân tích hoạt động của một chuỗi cửa hàng bán lẻ đa ngành, trong đó dữ liệu bao gồm thông tin chi tiết về khách hàng, sản phẩm, khu vực kinh doanh và các đơn hàng thực tế. Thông qua việc áp dụng các phương pháp khai phá dữ liệu (Data Mining) và phân tích dự báo, nhóm tập trung khám phá những khía cạnh quan trọng như: doanh thu theo thời gian, khu vực, danh mục sản phẩm; hành vi mua sắm của khách hàng; và dự đoán xu hướng doanh thu cũng như khả năng khách hàng quay lại mua hàng trong tương lai.

Chương I: Tổng quan về khai phá dữ liệu và giới thiệu đề tài

1 Giới thiệu Tổng quan về Khai phá Dữ liệu

1.1 Khái niệm khai phá dữ liệu là gì?

Khai phá dữ liệu là một lĩnh vực nhằm tự động khai thác những thông tin tri thức đang tiềm ẩn trong dữ liệu. Hay nói cách khác, "khai thác kiến thức từ dữ liệu".

Ví dụ sinh động: Quá trình tìm ra một lượng nhỏ vàng từ một lượng lớn nguyên liệu thô. Khai phá dữ liệu là một lĩnh vực phát triển bền vững, mang lại nhiều lợi ích, triển vọng, ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống

Các kỹ thuật được áp dụng dựa trên CSDL, học máy, trí tuệ nhân tạo, lý thuyết thông tin, xác suất thống kê và tính toán hiệu năng cao.

1.2 Quy trình Khám phá tri thức

Quá trình khám phá tri thức là một chuỗi lặp gồm các bước:

1. Data cleaning - Làm sạch dữ liệu (Loại bỏ nhiễu và dữ liệu không nhất quán)
2. Data integration - Tích hợp dữ liệu (Nơi nhiều nguồn dữ liệu có thể được kết hợp)
3. Data selection - Lựa chọn dữ liệu (Dữ liệu liên quan đến nhiệm vụ phân tích được truy xuất từ cơ sở dữ liệu)
4. Data transformation – Biến đổi dữ liệu (Chuẩn hoá và làm mịn dữ liệu để đưa dữ liệu về dạng phù hợp cho các kỹ thuật khai phá ở các bước sau)
5. Data mining - Khai phá dữ liệu (Quá trình cần thiết nơi các phương pháp được áp dụng để trích xuất các mẫu dữ liệu).
6. Pattern evaluation - Đánh giá mẫu (Xác định các mẫu thực sự thú vị thể hiện kiến thức dựa trên các thước đo, tiêu chí nhất định).
7. Knowledge presentation - Trình bày kiến thức một cách trực quan.

1.3 Giới thiệu về đề tài

Trong bối cảnh doanh nghiệp bán lẻ ngày càng cạnh tranh gay gắt, dữ liệu đang trở thành yếu tố cốt lõi giúp doanh nghiệp hiểu rõ khách hàng, tối ưu hoạt động kinh doanh và ra quyết định chiến lược. Ngày nay, mỗi giao dịch mua hàng, mỗi đơn đặt hàng, hay thậm chí là hành vi tìm kiếm và lựa chọn sản phẩm của khách hàng đều để lại "dấu vết dữ liệu" quan

trọng. Việc khai thác và phân tích các nguồn dữ liệu này không chỉ giúp doanh nghiệp nhìn lại quá khứ mà còn dự đoán và định hướng tương lai.

Bộ dữ liệu 3A Superstore được xây dựng nhằm mô phỏng hoạt động của một chuỗi cửa hàng bán lẻ đa ngành, cung cấp thông tin chi tiết về khách hàng, sản phẩm, đơn hàng, khu vực kinh doanh và các chỉ số doanh thu – lợi nhuận. Đây là một nguồn dữ liệu phong phú và đa chiều, rất phù hợp cho các bài toán phân tích kinh doanh, đánh giá hiệu quả hoạt động, dự báo xu hướng, và đặc biệt là nghiên cứu về hành vi khách hàng (customer behavior) trong lĩnh vực bán lẻ.

Dữ liệu bao gồm nhiều bảng liên kết với nhau như:

- **Bảng khách hàng (Customer):** Chứa thông tin định danh và đặc điểm của khách hàng.
- **Bảng đơn hàng (Orders):** Lưu trữ thông tin về từng giao dịch, thời gian mua hàng, giá trị đơn hang.
- **Bảng sản phẩm (Categories):** Cung cấp thông tin về danh mục hàng hóa, loại sản phẩm, tên sản phẩm
- **Bảng khu vực hoặc vùng miền (Region/Geography):** Giúp phân tích sự khác biệt doanh thu và hành vi tiêu dùng theo địa lý.
- **Bảng chi tiết bán hàng (Sales Details):** Kết nối dữ liệu giữa khách hàng, sản phẩm và thời gian để tính toán doanh số và lợi nhuận.

Chương II: Mục tiêu và bài toán đặt ra

2.1 Mục tiêu:

Trong bối cảnh thị trường bán lẻ ngày càng cạnh tranh gay gắt, việc ra quyết định kinh doanh dựa trên cảm tính không còn mang lại hiệu quả. Các doanh nghiệp hiện đại cần dựa vào phân tích dữ liệu (data analytics) để hiểu rõ khách hàng, sản phẩm và thị trường của mình. Từ đó, họ có thể tối ưu hóa chiến lược kinh doanh, marketing, chuỗi cung ứng và dịch vụ khách hàng.

Các mục tiêu chính gồm:

1. Phân tích tổng quan hoạt động kinh doanh của 3A Superstore:
    - Đánh giá doanh thu, lợi nhuận, số lượng đơn hàng theo thời gian (năm, quý, tháng).
    - Xác định xu hướng tăng trưởng hoặc suy giảm doanh số của doanh nghiệp.
2. Phân tích hành vi và giá trị khách hàng:

    - Khám phá thói quen mua hàng của khách hàng theo khu vực, danh mục sản phẩm, thời điểm.
    - Phân loại khách hàng theo giá trị (RFM: Recency, Frequency, Monetary) để tìm ra nhóm khách hàng trung thành và tiềm năng.
    - Đánh giá mức độ đóng góp của từng nhóm khách hàng vào tổng doanh thu.

3. Phân tích danh mục sản phẩm và khu vực kinh doanh:

    - Xác định danh mục sản phẩm hoặc khu vực mang lại doanh thu và lợi nhuận cao nhất.
    - Phát hiện các sản phẩm có doanh thu thấp hoặc bán chậm, từ đó đề xuất các chính sách khuyến mãi hoặc tái cấu trúc danh mục hàng hóa.
    - Đánh giá sự khác biệt trong hành vi tiêu dùng giữa các vùng miền.

4. Dự báo xu hướng doanh thu và hành vi tiêu dùng trong tương lai:

    - Sử dụng các mô hình phân tích và dự báo (như Prophet, ARIMA hoặc hồi quy tuyến tính) để dự đoán doanh thu trong các tháng tới.

2.2 Bài toán đề ra

1. Phân tích mô tả (Descriptive Analysis):

    - Tính toán và trực quan hóa các chỉ số cơ bản như doanh thu, số lượng đơn hàng, lợi nhuận trung bình, tần suất mua hàng.
    - So sánh doanh thu giữa các năm, tháng và khu vực.

2. Phân tích hành vi khách hàng (Customer Analysis):

    - Phân loại khách hàng theo giá trị mua hàng (nhóm VIP, nhóm tiềm năng, nhóm ít mua).
    - Tính toán chỉ số RFM (Recency, Frequency, Monetary) để đánh giá mức độ trung thành và giá trị của từng khách hàng.

3. Phân tích sản phẩm và danh mục (Product Analysis):

    - Xác định các sản phẩm bán chạy và sản phẩm kém hiệu quả.
    - Tìm hiểu mối quan hệ giữa giá bán, số lượng bán và tổng doanh thu.

4. Phân tích theo khu vực (Regional Analysis):

    - So sánh doanh thu giữa các vùng như Akdeniz, Ege, Marmara, Karadeniz, v.v.
    - Phát hiện khu vực có tiềm năng phát triển hoặc đang suy giảm để đề xuất hướng đầu tư, mở rộng hoặc cải thiện.

5. Phân tích dự báo (Forecasting): Dựa trên dữ liệu lịch sử, xây dựng mô hình dự báo doanh thu trong các tháng hoặc quý tiếp theo và hành vi của khách hàng

Chương III: Mô tả dữ liệu và tiền xử lý dữ liệu

3.1 Mô tả dữ liệu:

3.1.1 Giới thiệu chung:

Tên đầy đủ: 3A Superstore (Market Orders Data-CRM).

Mục đích: Dữ liệu giả lập (simulated) cho chuỗi bán lẻ "3A Superstore", dùng để phân tích CRM (Customer Relationship Management), bán hàng, phân tích khách hàng, marketing analytics.

3.1.2 Dung lượng:

Gồm có 5 file dữ liệu:

- Branches_ENG.csv

- Categories_ENG.csv
- Customers_ENG.csv
- Order_Details.csv
- Orders.csv

Định dạng: Dữ liệu ở dạng bảng (tabular) — có nhiều bảng kết nối (interconnected tables) theo mô tả "consists of 5 interconnected tables".

3.1.3 Các thuộc tính chính trong tập dữ liệu:

1. Order_Details (Chi tiết đơn hàng)

ORDERID – Mã đơn hàng

AMOUNT – Số lượng sản phẩm

UNITPRICE – Đơn giá

TOTALPRICE – Tổng tiền

2. Orders (Đơn hàng)

ORDERID – Mã đơn hàng

DATE_ – Ngày đặt hàng

USERID – Mã khách hàng

BRANCH_ID – Mã chi nhánh

TOTALBASKET – Tổng giá trị đơn hàng

3. Customers (Khách hàng)

USERID – Mã khách hàng

USERGENDER – Giới tính

USERBIRTHDATE – Ngày sinh

REGION, CITY, DISTRICT – Khu vực sinh sống

4. Branches (Chi nhánh)

BRANCH_ID – Mã chi nhánh

REGION – Khu vực

CITY – Thành phố

5. Categories (Sản phẩm)

ITEMID – Mã sản phẩm

BRAND – Thương hiệu

CATEGORY1, CATEGORY2 – Danh mục chính và phụ

ITEMNAME – Tên sản phẩm

3.1.4 Mô tả thống kê cơ bản:

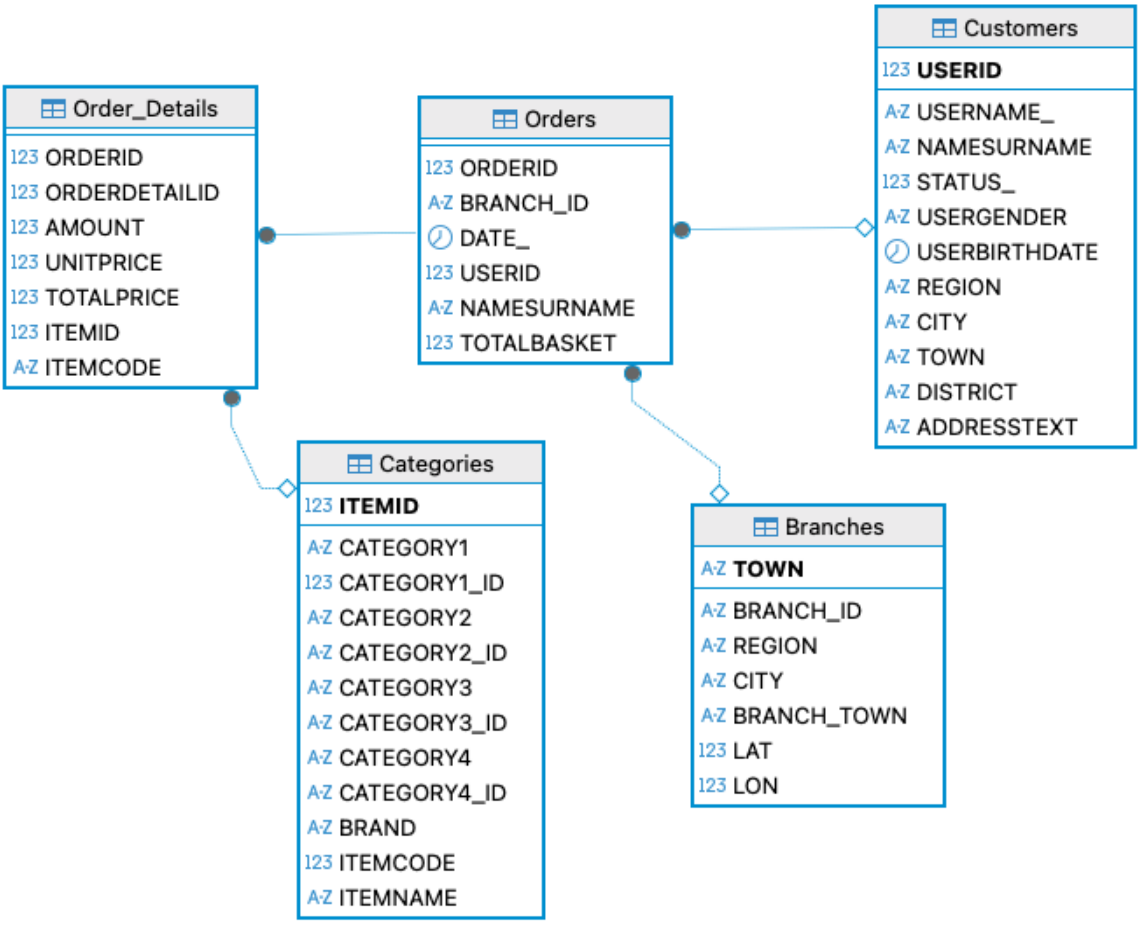
161 chi nhánh siêu thị khác nhau

27.000 mặt hàng siêu thị được phân loại

99.998 khách hàng và địa chỉ duy nhất

10.235.193 đơn hàng duy nhất

51.185.032 hàng chi tiết đơn hàng và 230.323.422 sản phẩm đã bán



Sơ đồ: Quan hệ thực thể

3.2 Làm sạch và tiền xử lý dữ liệu:

Định nghĩa: Làm sạch và tiền xử lý dữ liệu là giai đoạn chuẩn bị dữ liệu trước khi phân tích hoặc xây dựng mô hình, bao gồm các bước phát hiện, chỉnh sửa hoặc loại bỏ dữ liệu sai, thiếu, trùng lặp, không hợp lệ, và chuyển đổi dữ liệu về định dạng phù hợp.

Các bước làm sạch dữ liệu gồm:

- Xử lý giá trị bị thiếu
- Xử lý giá trị trùng lặp
- Chuyển đổi kiểu dữ liệu

- Xử lý giá trị ngoại lại (Outlier)
- Chuẩn hoá dữ liệu

3.2.1 Tạo một DataFrame hoàn chỉnh:

shape: (5, 9)

USERID	ORDERID	ITEMID	TOTALPRICE	…	CATEGORY1	ITEMNAME	REGION	DATE_
i64	i64	i64	str		str	str	str	datetime[μs]
64244	9370883	7563	40,46	…	Home	THE ART OF RAISING CHILDREN	Marmara	2022-07-20 00:00:00
38344	8136756	9892	49,04	…	Food	ÜLKER 204-7 BISKREM MINI KAK.9…	Ege	2023-06-30 00:00:00
58943	1228811	21604	135,600000000000002	…	Home	BOOK CRIME AND PUNISHMENT -NEW…	Akdeniz	2022-07-05 00:00:00
58943	1228811	15225	31,26	…	Tea-Coffee-Sugar	PALM BASIL LEAF 50 GR	Akdeniz	2022-07-05 00:00:00
58943	1228811	25934	57,22	…	Home	BIG SECRETS OF PEOPLE WHO HAVE…	Akdeniz	2022-07-05 00:00:00

✅ Tổng số dòng sau khi gộp: 51,184,395

Bảng: Bộ dữ liệu ban đầu

Kết quả: Hợp nhất dữ liệu từ nhiều bảng (chi tiết đơn hàng, đơn hàng, khách hàng, chi nhánh, danh mục sản phẩm) qua khoá nối tạo thành một DataFrame duy nhất giúp dễ dàng phân tích, thống kê hoặc huấn luyện mô hình.

3.2.2 Xử lý giá trị bị thiếu:

Định nghĩa: Xử lý giá trị thiếu là quá trình phát hiện và khắc phục các ô dữ liệu bị trống (NaN, Null, None, Missing) trong tập dữ liệu, nhằm đảm bảo dữ liệu đầy đủ và không gây sai lệch khi phân tích hoặc huấn luyện mô hình.

USERID	ORDERID	ITEMID	TOTALPRICE	…	CATEGORY1	ITEMNAME	REGION	DATE_
u32	u32	u32	u32		u32	u32	u32	u32
0	0	0	0	…	0	0	15891143	0

Sau khi kiểu tra dữ liệu ta thấy cột REGION đang có giá trị thiếu nên ta cần phải xử lý cột này bằng cách điền giá trị “Unknown”vào các dòng giá trị thiếu.

Kết quả: Giúp dữ liệu đầy đủ và đồng nhất, tránh lỗi khi tính toán hoặc huấn luyện mô hình và giúp giữ lại nhiều thông tin nhất có thể mà vẫn đảm báo độ chính xác.

3.2.3 Xử lý giá trị ngoại lai:

Định nghĩa: Xử lý giá trị ngoại lai là quá trình phát hiện và loại bỏ hoặc điều chỉnh các giá trị dữ liệu bất thường (quá lớn hoặc quá nhỏ so với phần còn lại) trong tập dữ liệu, nhằm tránh làm sai lệch kết quả phân tích và mô hình hóa.

Trước khi xử lý ngoại lai:

statistic	TOTALPRICE	TOTALBASKET
str	f64	f64
count	5.1184395e7	5.1184395e7
null_count	0.0	0.0
mean	256.045207	1621.578228
std	548.116603	1660.616185
min	0.0	0.0
25%	42.45	713.28
50%	116.06	1285.13
75%	288.12	2108.49
max	59520.96	307683.45

Giá trị tối đa cao bất thường so với 75%, nên khả năng có ngoại lai (outlier).

Sau khi xử lý ngoại lai:

statistic	TOTALPRICE	TOTALBASKET
str	f64	f64
count	4.9265163e7	4.9265163e7
null_count	0.0	0.0
mean	226.502109	1530.368414
std	293.922969	1102.581637
min	3.47	51.28
25%	44.58	722.02
50%	117.4	1277.71
75%	282.82	2062.06
max	2029.84	6393.53

Kết quả: Sau khi xử lý giá trị ngoại lai, dữ liệu trở nên ổn định và hợp lý hơn, không còn các giá trị bất thường gây sai lệch. Nhờ đó, kết quả phân tích và mô hình dự báo chính xác và đáng tin cậy hơn.

3.2.4 Chuyển đổi kiểu dữ liệu:

Chuyển đổi kiểu dữ liệu là quá trình thay đổi định dạng hoặc kiểu của dữ liệu (data type) từ dạng này sang dạng khác nhằm đảm bảo dữ liệu có thể xử lý, tính toán và phân tích đúng cách.

Trước chuyển đổi kiểu dữ liệu:

statistic	TOTALPRICE	TOTALBASKET
str	str	str
count	51184395	51184395
null_count	0	0
mean	null	null
std	null	null
min	0	0
25%	null	null
50%	null	null
75%	null	null
max	9999,960000000001	9999,789999999997

Sau khi chuyển đổi kiểu dữ liệu:

statistic	TOTALPRICE	TOTALBASKET
str	f64	f64
count	5.1184395e7	5.1184395e7
null_count	0.0	0.0
mean	256.045207	1621.578228
std	548.116603	1660.616185
min	0.0	0.0
25%	42.45	713.28
50%	116.06	1285.13
75%	288.12	2108.49
max	59520.96	307683.45

Kết quả: Có định dạng thống nhất và chính xác, dễ dàng xử lý, tính toán và trực quan hoá.

3.2.5 Chuẩn hoá dữ liệu:

Định nghĩa: quá trình đưa dữ liệu về dạng thống nhất và có cùng thang đo nhằm đảm bảo tính nhất quán, dễ so sánh và thuận tiện cho phân tích hoặc mô hình hoá.

Kết quả: một bộ dữ liệu có định dạng thống nhất, thang đo đồng đều và dễ sử dụng cho các bước phân tích hoặc mô hình hoá tiếp theo.

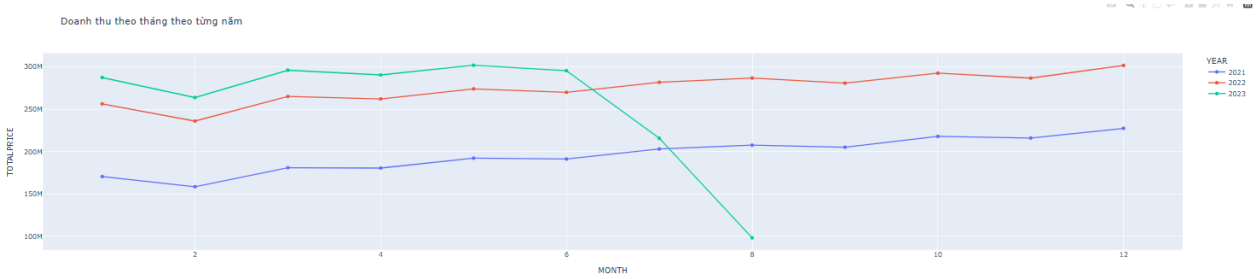
Chương IV: Phân tích và Phương pháp khai phá dữ liệu

4.1 Phân tích mô tả (Descriptive Analysis):

Phân tích mô tả là quá trình tổng hợp, trình bày và mô tả các đặc điểm cơ bản của dữ liệu thông qua các chỉ số thống kê (như trung bình, trung vị, độ lệch chuẩn, tần suất) và biểu đồ trực quan (như biểu đồ cột, đường, tròn, histogram...).

Mục tiêu của phân tích mô tả là hiểu rõ bức tranh tổng quan về dữ liệu, phát hiện các xu hướng, mô hình hoặc điểm bất thường, làm nền tảng cho các phân tích chuyên sâu hơn như phân tích dự đoán hoặc mô hình hóa.

4.1.1 Doanh thu theo năm:



Biểu đồ: Doanh thu theo từng năm

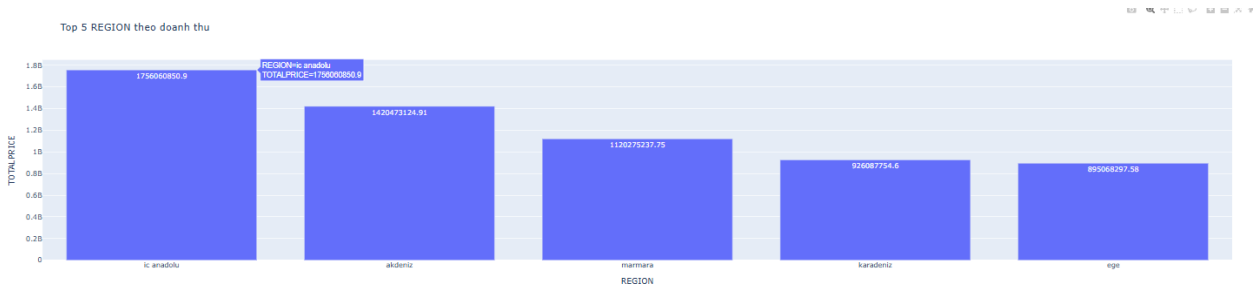
Năm 2021: ≈ 3.34 tỷ

Năm 2022: ≈ 4.90 tỷ

Năm 2023: ≈ 2.94 tỷ

Nhận xét: Năm 2022 đạt doanh thu cao nhất (~4.9 tỷ), cho thấy đây là giai đoạn kinh doanh hiệu quả nhất. Năm 2021 thấp hơn (~3.3 tỷ), nhưng vẫn giữ xu hướng tăng đều qua các tháng, cho thấy sự tăng trưởng ổn định. Năm 2023 lại có sự sụt giảm mạnh từ tháng 7 trở đi, có thể do thiếu dữ liệu các tháng cuối năm hoặc do hoạt động kinh doanh bị chững lại. Nhìn chung, doanh nghiệp đã có sự tăng trưởng tốt giai đoạn đầu nhưng cần xem xét nguyên nhân khiến doanh thu giảm trong năm 2023 để có biện pháp cải thiện.

4.1.2 Doanh thu theo khu vực



Biểu đồ: Doanh thu theo khu vực

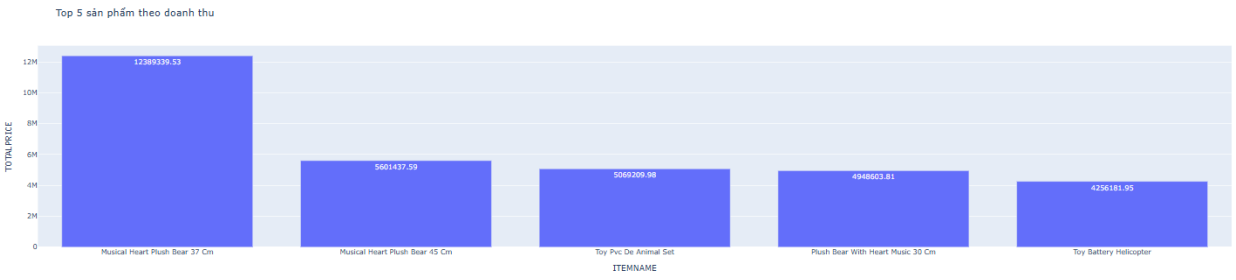
Vùng ic anadolu (hạng 1) đạt 1.76 tỷ đơn vị tiền tệ, cao hơn 23.6% so với vùng xếp thứ hai là akdeniz (1.42 tỷ).

Vùng akdeniz và Vùng marmara cùng thuộc nhóm có doanh thu trên 1 tỷ, tạo thành nhóm 3 vùng chủ lực đóng góp vào tổng doanh thu.

Vùng karadeniz và Vùng ege có doanh thu tương đương nhau (khoảng 0.9 tỷ) và chỉ bằng khoảng một nửa (51% - 53%) so với vùng dẫn đầu.

Nhận xét: ic anadolu dẫn đầu tuyệt đối với 1.76 tỷ doanh thu. akdeniz (1.42 tỷ) và marmara (1.12 tỷ) là nhóm trọng điểm trên 1 tỷ. karadeniz (0.93 tỷ) và ege (0.90 tỷ) ở nhóm dưới, chỉ đạt khoảng 51% doanh thu của vùng dẫn đầu, cho thấy sự phân hóa mạnh mẽ giữa các vùng.

4.1.3 Doanh thu theo sản phẩm

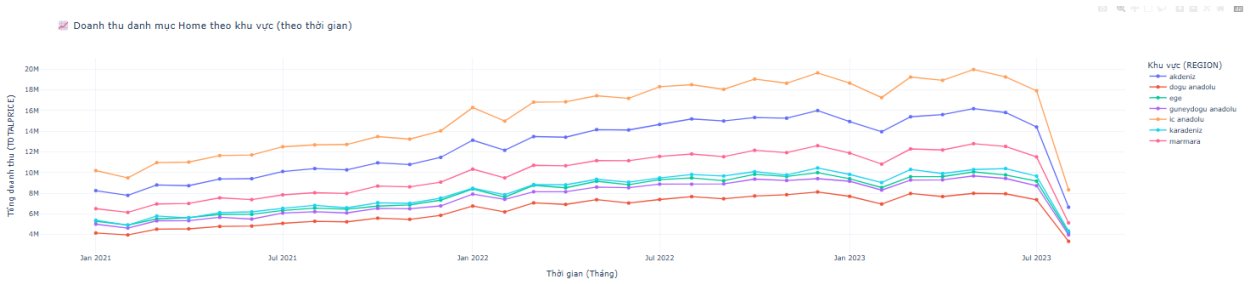


Biểu đồ: Doanh thu theo sản phẩm

Sự Thống Trị Tuyệt Đối: Sản phẩm Musical Heart Plush Bear 37 Cm là mặt hàng chủ lực áp đảo với doanh thu 12.39 triệu, chiếm 38.40% tổng doanh thu của Top 5. Sản phẩm này gấp 2.21 lần doanh thu của mặt hàng xếp thứ hai. Nhóm Hỗ Trợ: Bốn sản phẩm còn lại (5.60 triệu đến 4.26 triệu) có doanh thu tương đối đồng đều, hoạt động như một nhóm hỗ trợ. Tổng doanh thu của nhóm này (19.88 triệu) chỉ lớn hơn một chút so với doanh thu của riêng sản phẩm dẫn đầu.

Nhận xét: Sản phẩm Musical Heart Plush Bear 37 Cm là nguồn doanh thu chủ lực với 12.39 triệu, áp đảo toàn bộ nhóm và gấp 2.21 lần sản phẩm đứng thứ hai. Bốn sản phẩm còn lại thuộc nhóm bổ sung, có doanh thu dao động hẹp từ 4.26 triệu đến 5.60 triệu. Sự tập trung doanh thu cao độ vào một mặt hàng tạo ra rủi ro lớn và cần chiến lược đa dạng hóa.

4.1.4 Doanh thu của danh mục Home theo khu vực:



Biểu đồ: Doanh thu danh mục Home theo khu vực

Xu hướng chính: Tổng doanh thu của tất cả các khu vực đều tăng trưởng liên tục từ đầu năm 2021 đến giữa năm 2023.

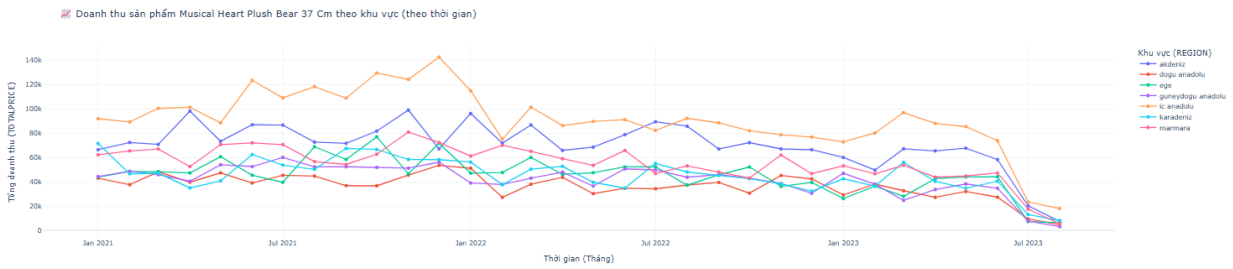
Dẫn đầu: Khu vực Marmara và Ic Anadolu duy trì mức doanh thu cao vượt trội so với các khu vực còn lại trong suốt giai đoạn.

Nhận xét: Xu hướng tăng trưởng: Nhìn chung, tổng doanh thu của tất cả các khu vực đều có xu hướng tăng ổn định trong giai đoạn từ đầu năm 2021 đến đầu năm 2023.

Sụt giảm cuối kỳ: Doanh thu của tất cả các khu vực đều giảm mạnh và đột ngột vào giai đoạn cuối cùng của biểu đồ (khoảng giữa đến cuối năm 2023).

Phân hóa khu vực: Các khu vực có sự phân hóa rõ rệt về quy mô doanh thu, duy trì một trật tự tương đối ổn định trong suốt thời gian.

4.1.5 Doanh thu của sản phẩm Musical Heart Plush Bear 37 Cm theo khu vực:



Biểu đồ: Doanh thu sản phẩm Musical Heart Plush Bear 37 Cm theo khu vực

Doanh thu sản phẩm Musical Heart Plush Bear 37 Cm biến động theo thời gian, đạt mức cao trong giai đoạn 2021–2022, đặc biệt tại Akdeniz và Marmara. Tuy nhiên, từ năm 2023, doanh thu giảm rõ rệt ở tất cả khu vực, thể hiện xu hướng suy giảm nhu cầu hoặc thị trường bão hòa.

Nhận xét: Sản phẩm đang mất dần sức hút, cần xem xét lại chiến lược kinh doanh và quảng bá. Doanh nghiệp nên tập trung vào các khu vực tiêu thụ mạnh như Akdeniz và Marmara, đồng thời tìm biện pháp kích cầu và làm mới sản phẩm.

4.2 Dự đoán doanh thu sản phẩm/danh mục theo khu vực:

Dự đoán sử dụng phương pháp khai phá là Hồi quy (Regression). Kỹ thuật hồi quy được sử dụng để dự đoán một giá trị liên tục (continuous value) dựa trên các biến đầu vào. Đối với dự án này, chúng tôi tập trung vào Hồi quy để dự đoán TOTALPRICE.

XGBRegressor là mô hình hồi quy dựa trên thuật toán Extreme Gradient Boosting, kết hợp nhiều cây quyết định (Decision Trees) nhỏ để tạo ra một mô hình dự đoán mạnh mẽ và chính xác.

Ưu điểm: chính xác, nhanh, nhiều tuỳ chỉnh, xử lý missing.

Nhược điểm: cần tuning, xử lý categorical trước, interpretability cần công cụ bổ sung (SHAP).

Các tham số quan trọng:

- n_estimators: Số cây quyết định được huấn luyện.
- learning_rate: Tốc độ học (đóng góp mỗi cây).
- max_depth: Độ sâu tối đa của mỗi cây.
- subsample: Tỷ lệ mẫu ngẫu nhiên mỗi cây.
- colsample_bytree: Tỷ lệ cột ngẫu nhiên mỗi cây.
- reg_lambda: Hệ số phạt L2.
- reg_alpha: Hệ số phạt L1.
- random_state: Hạt giống ngẫu nhiên.
- objective: Hàm mục tiêu hồi quy (MSE).
- eval_set: Bộ dữ liệu kiểm định (validation).

4.3 Dự đoán xác suất khách hang mua trong 3 tháng tới

Sử dụng mô hình Logistic Regression với bộ đặc trưng:

- RFM: Recency (số ngày từ lần mua gần nhất), Frequency (số đơn hàng), Monetary (tổng giá trị mua), AvgBasketSize (giá trung bình mỗi sản phẩm), NumCategories (số danh mục sản phẩm). Thông tin phân cụm khách hàng (Cluster_0, Cluster_1, …) được one-hot encode. Dữ liệu được chuẩn hóa bằng StandardScaler trước khi huấn luyện.

- Train/Test Split theo tỉ lệ 70/30, stratified theo nhãn PurchaseNext90 để giữ tỉ lệ khách mua/không mua.

- Mô hình huấn luyện với tham số class_weight='balanced' để cân bằng nhãn.

Ưu điểm: Mô hình đơn giản, dễ hiểu và có khả năng giải thích trực tiếp mối quan hệ giữa các đặc trưng RFM và phân cụm khách hàng với xác suất mua. Logistic Regression phù hợp với bài toán nhị phân và có chi phí tính toán thấp, giúp quá trình huấn luyện diễn ra nhanh chóng. Sử dụng tham số class_weight='balanced' để xử lý các vấn đề mất cân bằng nhãn, trong khi chuẩn hóa dữ liệu bằng StandardScaler giúp mô hình hội tụ ổn định

Nhược điểm: Logistic Regression không thể mô hình hóa tốt các tương tác phi tuyến giữa các biến, do đó hiệu quả dự đoán có thể giảm nếu mối quan hệ giữa các đặc trưng RFM và xác suất mua không tuyến tính. Ngoài ra, mô hình vẫn có thể bị ảnh hưởng bởi các giá trị ngoại lai, mặc dù dữ liệu đã được xử lý winsorize trước đó

4.4 Dự đoán giá trị vòng đời khách hàng 3 tháng tới

Sử dụng mô hình LightGBM Regressor với cùng bộ đặc trưng như mô hình Purchase. Giá trị mục tiêu (CLV90_log) được log-transform để giảm ảnh hưởng của các giá trị cực đoan. Giá trị CLV được cắt bớt các outliers trên 99% để ổn định mô hình (CLV90_capped).

Trong huấn luyện, sử dụng trọng số mẫu (sample_weight) nhằm giảm bias từ khách hàng có giá trị CLV quá lớn.

Ưu điểm: LightGBM có khả năng xử lý tốt các dữ liệu phi tuyến và tự động phát hiện tương tác giữa các đặc trưng, giúp dự đoán CLV chính xác hơn. Mô hình hỗ trợ sử dụng trọng số mẫu, từ đó giảm thiểu bias với những khách hàng có giá trị CLV cực đoan. Việc áp dụng log-transform và cắt bớt các giá trị ngoại lai (CLV90_capped) giúp LightGBM ổn định hơn trong quá trình huấn luyện. Ngoài ra, LightGBM có khả năng mở rộng tốt và dự đoán nhanh trên các tập dữ liệu lớn

Nhược điểm: khó giải thích trực quan so với mô hình tuyến tính, việc tinh chỉnh hyperparameters là cần thiết để đạt hiệu quả tối ưu, và nếu dữ liệu quá nhỏ hoặc quá ít khách hàng có CLV lớn, mô hình vẫn có thể học lệch mặc dù đã sử dụng trọng số mẫu.

4.5 Dự đoán khả năng khách hàng rời bỏ 3 tháng tới

Sử dụng mô hình RandomForestClassifier với bộ đặc trưng RFM và cluster tương tự. Train/Test Split 70/30, stratified theo nhãn Churn90. Mô hình cân bằng nhãn bằng class_weight='balanced' để xử lý khách hàng mua ít hoặc không mua.

Ưu điểm: Random Forest mạnh mẽ với dữ liệu phi tuyến và nhiều chiều, tự động xử lý các tương tác giữa các biến. Việc sử dụng class_weight='balanced' giúp cân bằng nhãn, đồng thời mô hình ít bị overfitting hơn so với cây quyết định đơn lẻ. Ngoài ra, Random Forest cho phép tính toán feature importance, giúp đánh giá mức độ đóng góp của từng đặc trưng vào dự đoán

Nhược điểm: xác suất dự đoán đôi khi không mượt và mang tính "bậc thang" do đặc trưng ensemble, mô hình tốn bộ nhớ và thời gian tính toán hơn Logistic Regression, đặc biệt khi số lượng cây và độ sâu lớn, đồng thời việc giải thích trực quan khó khăn hơn so với các mô hình tuyến tính.

Chương V: Kết quả và đánh giá mô hình.

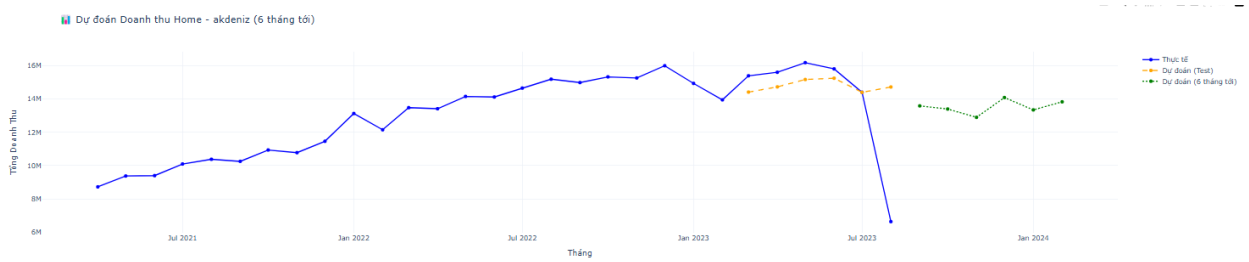
5.1 Kết quả dự đoán doanh thu danh mục theo khu vực.

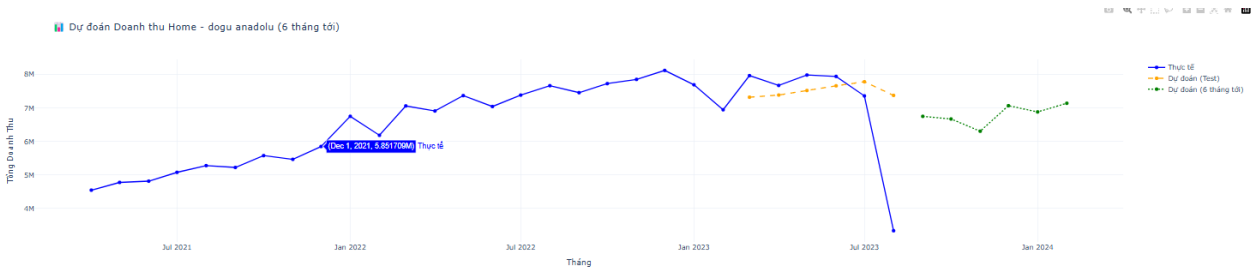
Độ chính xác:

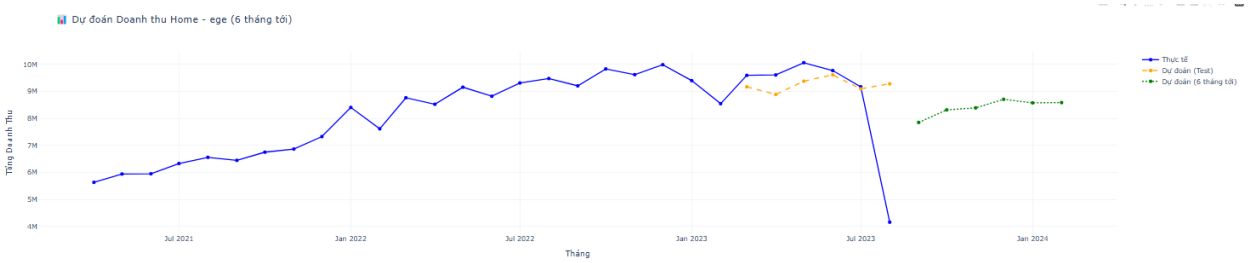
📈 Độ chính xác mô hình XGBoost (Test giai đoạn gần nhất):

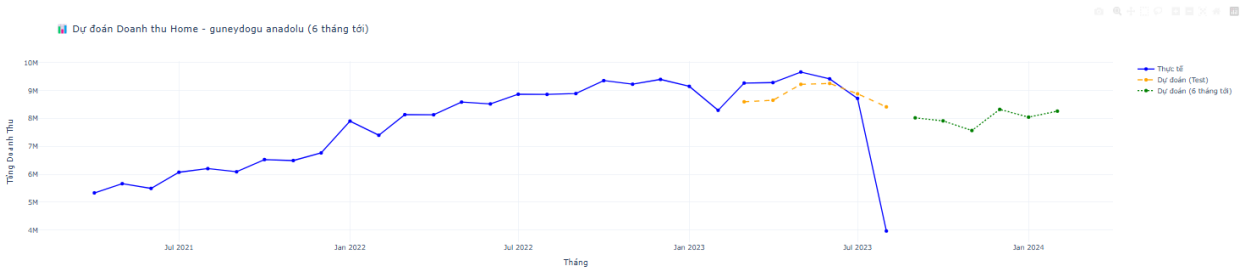
	REGION	MAE	RMSE	MAPE (%)
3	guneydogu anadolu	1087947.000000	1867335.010000	22.410000
4	ic anadolu	2294804.210000	4178751.570000	23.330000
0	akdeniz	1918381.030000	3373645.260000	23.910000
2	ege	1197390.520000	2136101.830000	24.030000
5	karadeniz	1274621.890000	2196431.220000	24.160000
1	dogu anadolu	1021808.940000	1696230.680000	24.620000
6	marmara	1618161.490000	2662955.060000	25.130000

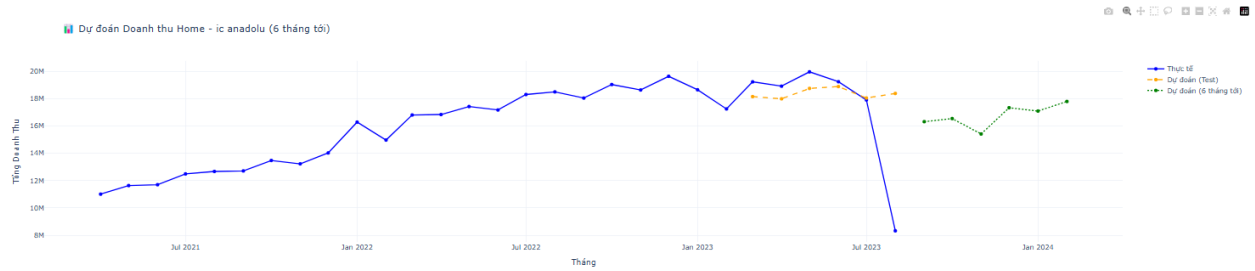
Dự đoán theo từng khu vực:

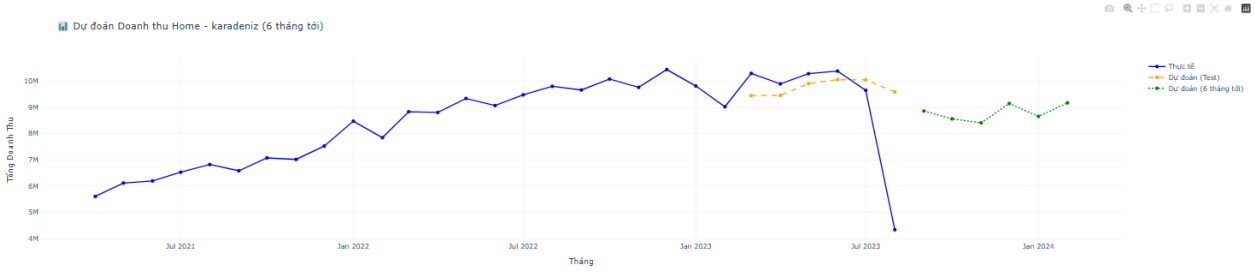


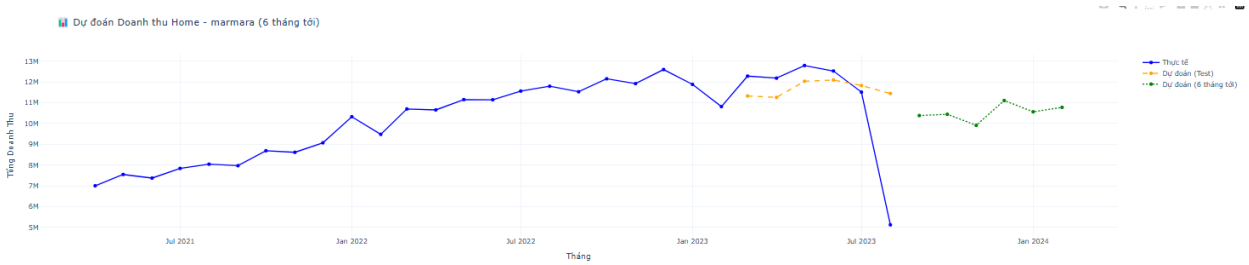

Dự đoán Doanh thu Home - dogu anadolu (6 tháng tới)
Tổng Doanh Thu
Tháng
Thực tế
Dự đoán (Test)
Dự đoán (6 tháng tới)
(Dec 1, 2021, 5.881709M) Thực tế
8M
7M
6M
5M
4M
Jul 2021
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Jan 2024


Dự đoán Doanh thu Home - ege (6 tháng tới)
Tổng Doanh Thu
Tháng
Thực tế
Dự đoán (Test)
Dự đoán (6 tháng tới)
10M
9M
8M
7M
6M
5M
4M
Jul 2021
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Jan 2024


Dự đoán Doanh thu Home - guneydogu anadolu (6 tháng tới)
Tổng Doanh Thu
Tháng
Thực tế
Dự đoán (Test)
Dự đoán (6 tháng tới)
10M
9M
8M
7M
6M
5M
4M
Jul 2021
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Jan 2024


Dự đoán Doanh thu Home - ic anadolu (6 tháng tới)
Tổng Doanh Thu
Tháng
Thực tế
Dự đoán (Test)
Dự đoán (6 tháng tới)
20M
18M
16M
14M
12M
10M
8M
Jul 2021
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Jan 2024


Dự đoán Doanh thu Home - karadeniz (6 tháng tới)
Tổng Doanh Thu
Tháng
Thực tế
Dự đoán (Test)
Dự đoán (6 tháng tới)
10M
9M
8M
7M
6M
5M
4M
Jul 2021
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Jan 2024



Biểu đồ: Dự đoán doanh thu danh mục Home theo từng khu vực

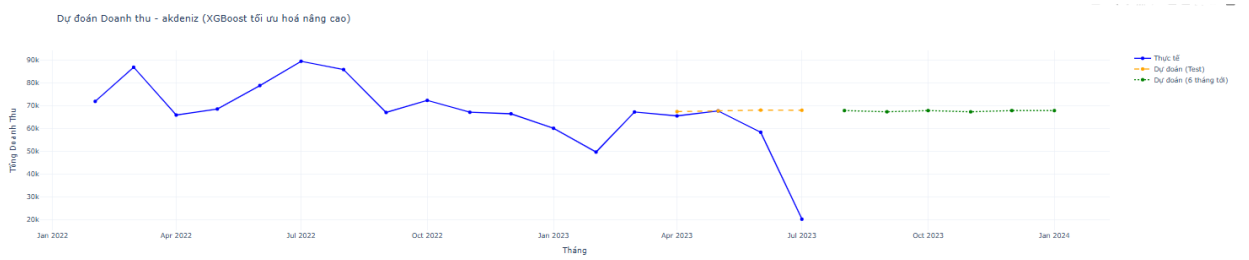
5.2 Kết quả dự đoán doanh thu sản phẩm theo khu vực:

Độ chính xác:

Độ chính xác mô hình XGBoost:

	REGION	MAE	RMSE	MAPE (%)
6	marmara	5360.400000	6088.840000	11.840000
5	karadeniz	6379.470000	7985.200000	14.200000
2	ege	6194.070000	6822.420000	17.710000
1	dogu anadolu	6042.670000	6815.370000	21.320000
3	guneydogu anadolu	7894.750000	9205.880000	27.350000
0	akdeniz	14850.960000	24393.110000	64.100000
4	ic anadolu	34431.590000	42722.260000	134.340000

Dự đoán theo khu vực:



Biểu đồ: Dự đoán doanh thu sản phẩm Musical Heart Plush Bear 37 Cm theo từng khu vực

5.3 Kết quả của dự đoán hành vi khách hàng.

- Độ chính xác khi dự đoán tất cả khách hàng:

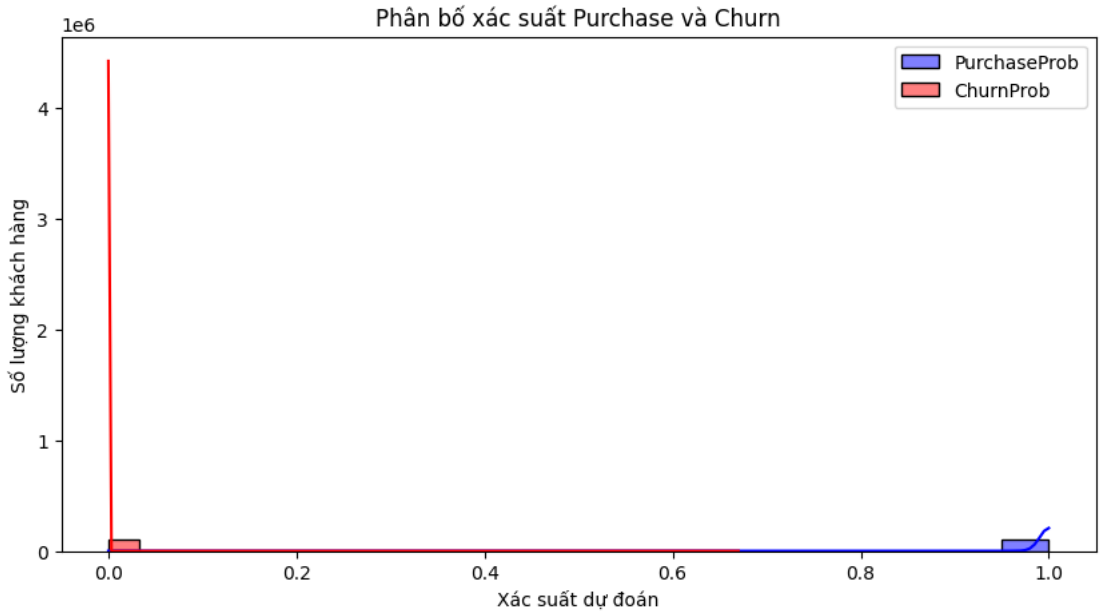
```
Purchase Accuracy: 98.93%
Churn Accuracy: 100.00%
CLV MAPE: 51.94%, SMAPE: 42.20%, RMSE: 7069.49
```

- Độ chính xác khi dự đoán một khách hàng bất kì:

```
📊 KẾT QUẢ DỰ ĐOÁN CHO USER 80281 (90 ngày tới)
🛍️ PurchaseNext90: Prob=100.00% | Thực tế=1, Accuracy=✅
💰 CLV90: Dự đoán=10,583 | Thực tế=11,710
   ├─ MAPE=9.62% | SMAPE=10.11% | RMSE=1126.88
   └─ (log-scale) MAPE=1.08% | RMSE=0.10
⚠️ Churn90: Prob=0.00% | Thực tế=0, Accuracy=✅
```

5.3.1 Kết quả dự đoán xác suất khách hàng có tiếp tục mua trong ba tháng tới.

- Kết quả khi dự đoán tất cả khách hàng:



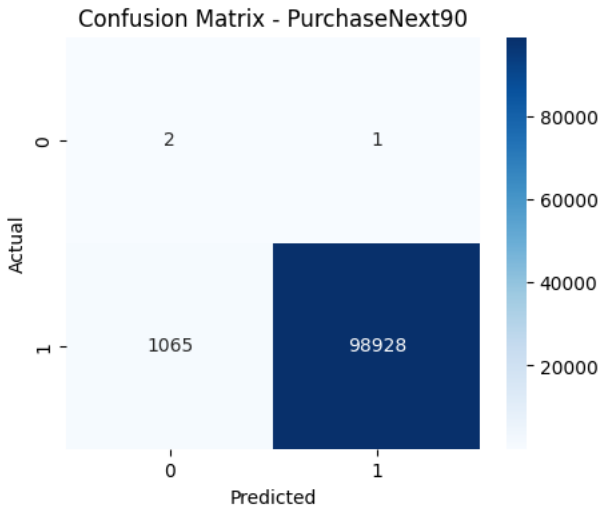
Biểu đồ: Dự đoán khách hàng hành vi của tất cả khách hàng.

Nhận xét:

- Đa phần khách hàng có xác suất PurchaseProb gần 1, tức nhiều khách hàng có khả năng cao sẽ mua.
- Đa phần khách hàng có xác suất ChurnProb gần 0**,** tức hầu hết khách hàng ít khả năng rời bỏ.

- Có một số nhỏ khách hàng nằm ở giữa (0.5–0.7), nhưng rất ít so với tổng số lượng (trục y cực cao ở 0 và gần 1).
- Biểu đồ này cũng chỉ ra mô hình gần như phân tách hoàn toàn hai nhóm, mô hình đang overfitting vì số lượng dự đoán quá cực đoan.

- Ma trận nhầm lẫn:

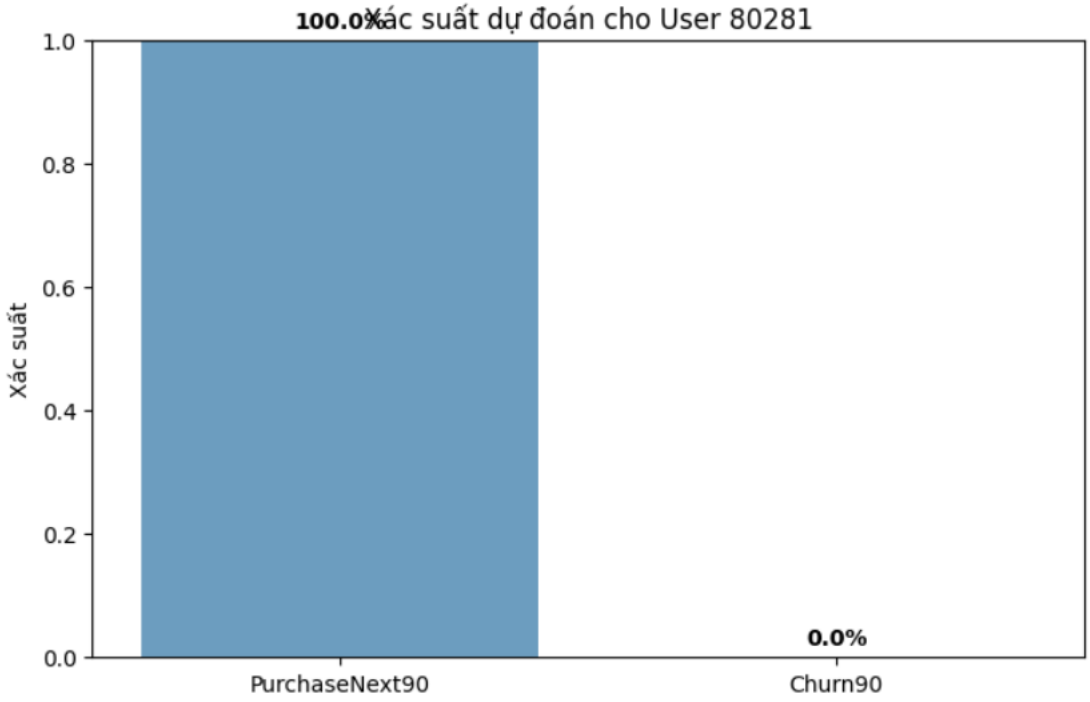


Ma trận nhầm lẫn: Dự đoán hành vi khách hàng.

Nhận xét:

- Mô hình dự đoán rất tốt cho lớp chiếm đa số (lớp 1) nhưng không học được gì cho lớp ít (lớp 0).
- Dữ liệu đang mất cân bằng: Accuracy cao nhưng mô hình thiếu khả năng phân biệt lớp nhỏ.
- Cần thêm các biện pháp cân bằng dữ liệu (undersampling lớp lớn, oversampling lớp nhỏ, SMOTE…) hoặc sử dụng các metric khác như F1-score theo lớp, ROC-AUC, thay vì chỉ dựa vào accuracy.

- Kết quả khi dự đoán một khách hàng bất kì:



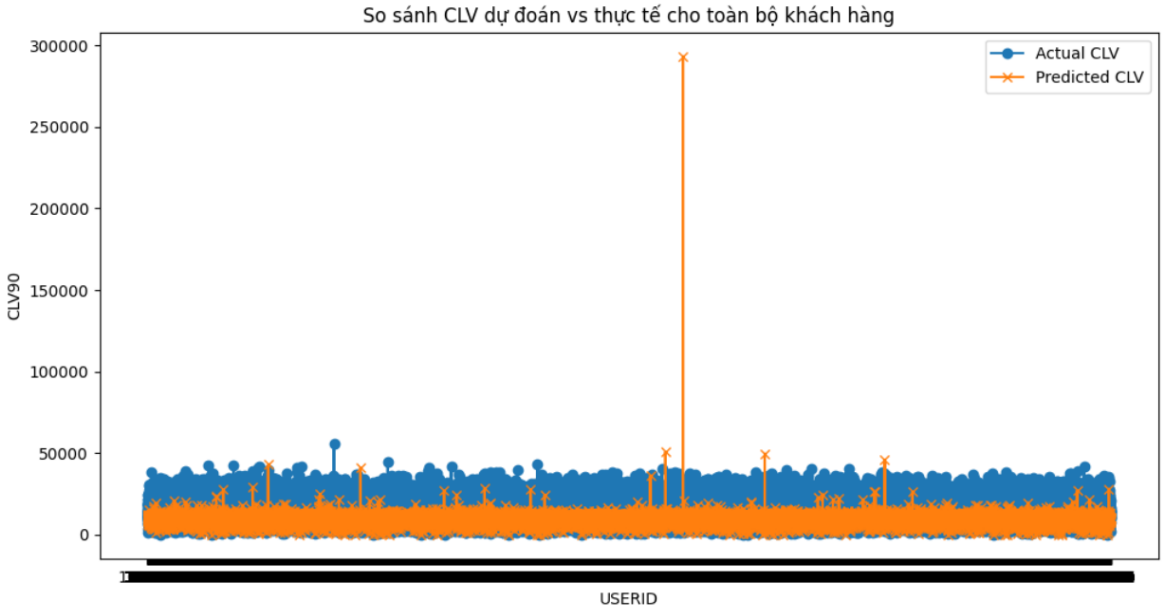
Biểu đồ: Dự đoán hành vi của một khách hàng bất kì.

Nhận xét:

- Đây là dự đoán cực kỳ chắc chắn (xác suất tuyệt đối), điều này có thể do mẫu dữ liệu của người dùng này có hành vi rất rõ ràng hoặc mô hình đang "quá tự tin".
- Nếu nhiều người dùng có xác suất 0% hoặc 100% như thế này, có thể mô hình đang overfitting

5.3.2 Kết quả dự đoán giá trị vòng đời của khách hàng trong ba tháng tới.

- Kết quả khi dự đoán toàn bộ khách hàng:

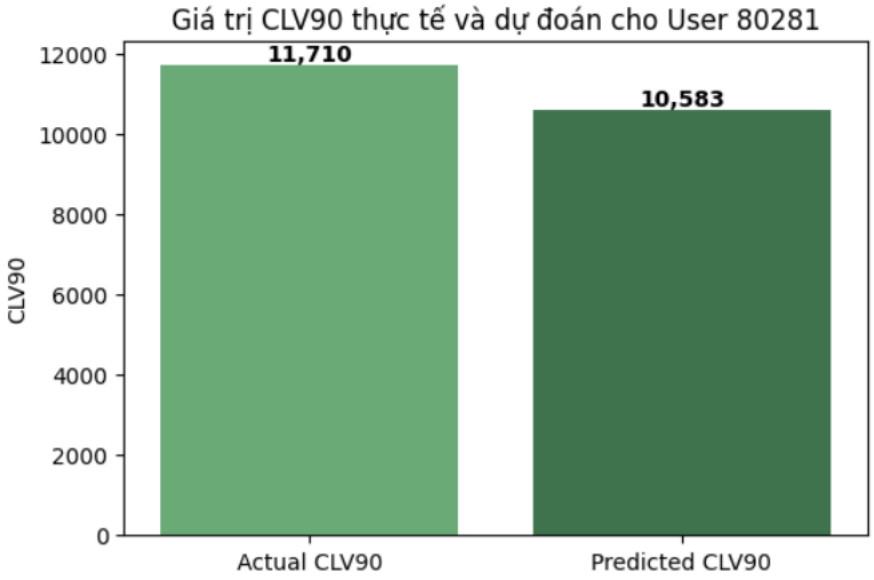


Biểu đồ: Dự đoán giá trị vòng đời của khách hàng.

Nhận xét:

- Giá trị thực tế (Actual CLV) nằm phân bố tập trung quanh mức từ 0 đến ~50,000.
- Giá trị dự đoán (Predicted CLV) nhìn chung thấp hơn so với CLV thực tế, với nhiều điểm nằm gần 0, điều này cho thấy mô hình có xu hướng underestimation (dự đoán thấp hơn thực tế).
- Có một số giá trị dự đoán cực kỳ cao (ví dụ ~300,000), rõ ràng là dự đoán sai lệch quá mức, có thể do mô hình bị ảnh hưởng bởi các điểm ngoại lai hoặc phân phối CLV quá lệch phải.
- Một số giá trị thực tế cũng rất cao, nhưng mô hình không dự đoán được chính xác, chứng tỏ khả năng xử lý các khách hàng "high-value" còn kém.
- Phần lớn các dự đoán nằm dưới giá trị thực tế, chứng tỏ mô hình chưa capture tốt những khách hàng có giá trị cao.
- Với phần lớn khách hàng có CLV thấp đến trung bình, mô hình dự đoán khá sát.

- Kết quả khi dự đoán một khách hàng bất kì:



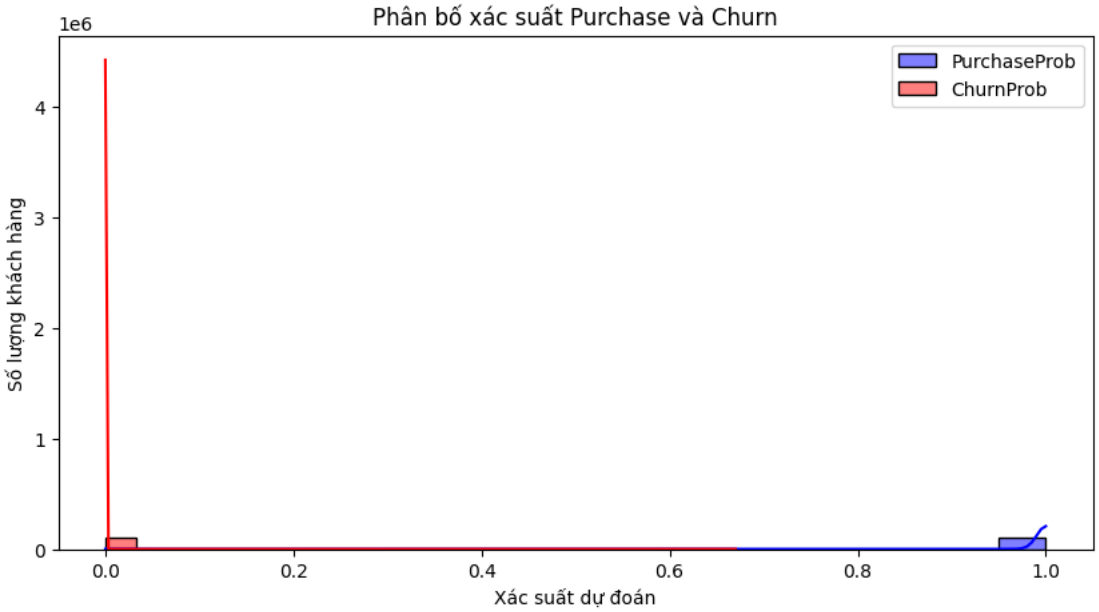
Biểu đồ: Dự đoán giá trị vòng đời một khách hàng bất kì.

Nhận xét:

- Giá trị CLV90 thực tế của User 80281 là **11,710** và dự đoán là **10,583**.
- Mô hình dự đoán thấp hơn giá trị thực khoảng **1,127**, tức lệch khoảng **~9.6%** so với giá trị thực.
- Mô hình dự đoán gần sát với giá trị thực, chỉ sai lệch một chút. Đây là mức chấp nhận được nếu mục tiêu là dự đoán tương đối. Quan sát cho thấy mô hình có xu hướng **dự đoán hơi thấp hơn thực tế** cho User này.

5.3.3 Kết quả dự đoán khả năng khách hàng rời bỏ trong ba tháng tới.

- Kết quả khi dự đoán tất cả khách hang bất kì:

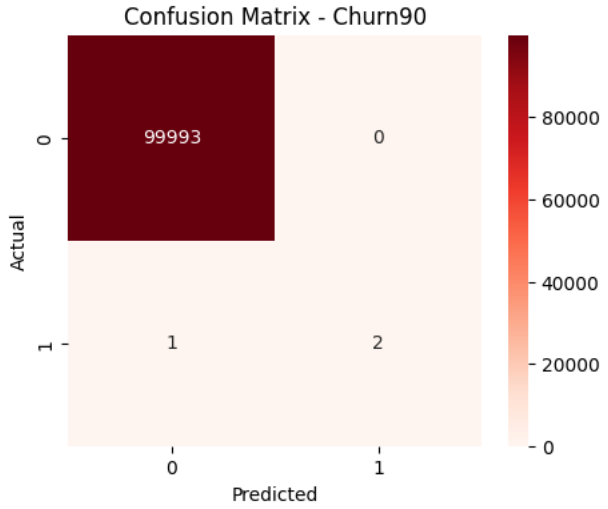


Biểu đồ: Dự đoán khả năng khách hàng rời bỏ.

Nhận xét:

- Đa phần khách hàng có xác suất PurchaseProb gần 1, tức nhiều khách hàng có khả năng cao sẽ mua.
- Đa phần khách hàng có xác suất ChurnProb gần 0**,** tức hầu hết khách hàng ít khả năng rời bỏ.
- Có một số nhỏ khách hàng nằm ở giữa (0.5–0.7), nhưng rất ít so với tổng số lượng (trục y cực cao ở 0 và gần 1).
- Biểu đồ này cũng chỉ ra mô hình gần như phân tách hoàn toàn hai nhóm, mô hình đang overfitting vì số lượng dự đoán quá cực đoan.

- Ma trận nhầm lẫn:

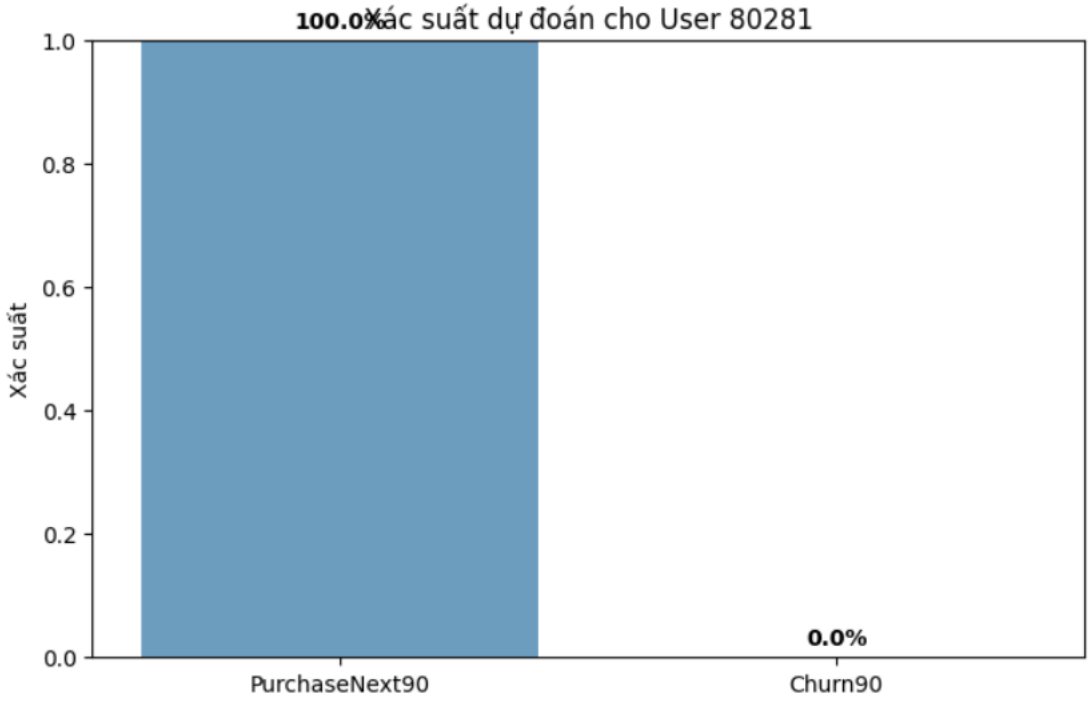


Ma trận nhầm lẫn: Dự đoán khả năng khách hàng rời bỏ.

Nhận xét:

- Actual 0 (khách không churn): 99,993 được dự đoán đúng là 0, 0 dự đoán sai là 1 → khả năng dự đoán khách không churn rất tốt, gần như hoàn hảo.
- Actual 1 (khách churn)**:** 1 được dự đoán là 0, 2 dự đoán đúng là 1 → số lượng khách churn rất ít và mô hình chỉ bắt đúng 2/3 trường hợp, còn 1 trường hợp nhầm thành không churn.
- Mô hình có độ chính xác tổng thể cực cao (99,995%), nhưng đó có thể là do dữ liệu mất cân bằng (khách churn quá ít so với không churn).
- Khả năng phát hiện churn (recall của lớp 1) còn hạn chế**:** chỉ 2/3 = ~66.7%.
- Nên xem xét cân bằng dữ liệu, hoặc dùng precision-recall**,** F1-score thay vì chỉ dựa vào accuracy.

- Kết quả khi dự đoán một khách hàng bất kì:



Biểu đồ: Dự đoán khả năng rời bỏ của một khách hàng bất kì.

Nhận xét:

- Đây là dự đoán cực kỳ chắc chắn (xác suất tuyệt đối), điều này có thể do mẫu dữ liệu của người dùng này có hành vi rất rõ ràng hoặc mô hình đang "quá tự tin".
- Nếu nhiều người dùng có xác suất 0% hoặc 100% như thế này, có thể mô hình đang overfitting

Chương VI: Kết luận và hướng phát triển.

6.1 Kết luận:

Qua quá trình thực hiện đề tài **"3A Superstore (Market Orders Data – CRM)"**, nhóm đã tiến hành toàn bộ quy trình khai phá dữ liệu từ khâu làm sạch, tiền xử lý, phân tích mô tả đến xây dựng mô hình dự đoán. Dữ liệu được hợp nhất từ nhiều bảng (Orders, Order Details, Customers, Branches, Categories), giúp hình thành một hệ thống thông tin hoàn chỉnh phản ánh hoạt động kinh doanh của chuỗi siêu thị bán lẻ.

Kết quả thu được cho thấy mô hình có độ chính xác tương đối và có thể áp dụng trong việc hỗ trợ ra quyết định kinh doanh. Nhìn chung, đề tài đã hoàn thành tốt các mục tiêu đề ra,

minh chứng được vai trò quan trọng của khai phá dữ liệu trong lĩnh vực bán lẻ – nơi mỗi quyết định kinh doanh đều có thể được dẫn dắt bởi những phân tích dữ liệu cụ thể và khoa học.

6.2 Hướng phát triển:

Trong tương lai, nhóm đề xuất mở rộng và hoàn thiện đề tài theo các hướng sau:

Mở rộng phạm vi dữ liệu:

- Thu thập thêm dữ liệu thực tế từ nhiều năm hơn, bao gồm các yếu tố mùa vụ, khuyến mãi, chi phí marketing để tăng độ chính xác của mô hình dự báo.
- Bổ sung dữ liệu về đánh giá sản phẩm, phản hồi khách hàng, kênh bán hàng (online/offline) để có góc nhìn toàn diện hơn.

Xây dựng hệ thống Dashboard trực quan:

- Phát triển bảng điều khiển (Dashboard) tương tác bằng Power BI hoặc Streamlit để hiển thị doanh thu, hành vi khách hàng, và kết quả dự báo theo thời gian thực.
- Hỗ trợ nhà quản lý dễ dàng theo dõi, so sánh và đưa ra quyết định dựa trên dữ liệu.