BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP. HỒ CHÍ MINH

KHOA KHOA CÔNG NGHỆ THÔNG TIN

----□🕮ᘐ----

BÁO CÁO MÔN HỌC

LẬP TRÌNH PHÂN TÍCH DỮ LIỆU 2

ĐỀ TÀI

DỰ ĐOÁN KHÁCH HÀNG TIỀM NĂNG MUA SẮM DỰA TRÊN DỮ LIỆU GIAO DỊCH VÀ HỒ SƠ KHÁCH HÀNG

Giảng viên hướng dẫn	**:** TS. LÊ TRỌNG NGỌC
Lớp học phần	**:** DHKHMT18A
Mã lớp học phần	**:** 420300233001
Nhóm	**:** Data Novices

TP.HCM, ngày 18 tháng 10 năm 2024

DANH SÁCH THÀNH VIÊN

STT	Họ và tên	MSSV	Ghi chú
1	Nguyễn Minh Phúc	22637001	
2	Phạm Gia Khánh	22724051	
3	Trần Minh Tú	22715731	
4	Nguyễn Đức Chiến	20066981	

LỜI CẢM ƠN

Trước tiên, nhóm em xin bày tỏ lòng biết ơn sâu sắc tới giảng viên hướng dẫn TS. Lê Trọng Ngọc, Khoa Công nghệ Thông tin, Trường Đại học Công nghiệp Thành phố Hồ Chí Minh người đã tận tình hướng dẫn, cung cấp kiến thức, tài liệu và phương pháp, giúp đỡ cho nhóm em hoàn thành tốt đề tài nghiên cứu này.

Mặc dù nhóm em cũng đã cố gắng rất nhiều, nhưng do giới hạn về thời gian và kiến thức, đề tài của nhóm em sẽ còn nhiều thiếu sót, rất mong nhận được ý kiến đóng góp từ quý thầy cô. Sau khi hoàn thành xong đề tài nghiên cứu này, nhóm em cũng đã học được rất nhiều kiến thức mới và bổ ích, là tiền đề cho sau này để nhóm em hoàn thành các nghiên cứu khác.

Nhóm em xin trân trọng cảm ơn!

MỤC LỤC

PHẦN MỞ ĐẦU

1. LÝ DO CHỌN ĐỀ TÀI

Trong thời đại số hóa và sự phát triển mạnh mẽ của thương mại điện tử, việc thấu hiểu hành vi khách hàng và dự đoán những khách hàng tiềm năng đã trở thành một trong những yếu tố quan trọng giúp doanh nghiệp tối ưu hóa chiến lược kinh doanh. Dự đoán chính xác khách hàng có khả năng mua sắm không chỉ giúp tiết kiệm chi phí tiếp thị, mà còn tạo điều kiện cho doanh nghiệp cung cấp các sản phẩm, dịch vụ phù hợp và kịp thời, từ đó nâng cao trải nghiệm của khách hàng.

Dữ liệu giao dịch và hồ sơ khách hàng là một nguồn tài nguyên quý giá, chứa đựng nhiều thông tin về thói quen, sở thích, và khả năng mua sắm của khách hàng. Việc ứng dụng các phương pháp học máy (Machine Learning) trong phân tích dữ liệu này mở ra khả năng dự đoán hành vi mua sắm với độ chính xác cao, giúp doanh nghiệp dễ dàng nhận diện và tiếp cận đúng đối tượng khách hàng.

Do đó, đề tài "Dự đoán khách hàng tiềm năng mua sắm dựa trên dữ liệu giao dịch và hồ sơ khách hàng" được chọn với mong muốn nhóm có thể ứng dụng các kỹ thuật phân tích dữ liệu tiên tiến nhằm giải quyết bài toán kinh doanh thực tiễn. Thông qua đề tài này, nhóm mong muốn xây dựng một mô hình dự đoán có hiệu quả, đồng thời cung cấp các phân tích chuyên sâu giúp doanh nghiệp có cái nhìn rõ ràng hơn về khách hàng của mình, từ đó cải thiện hiệu quả kinh doanh.

2. MỤC TIÊU

Phát triển một mô hình dự đoán khách hàng tiềm năng mua sắm dựa trên dữ liệu giao dịch và hồ sơ khách hàng. Mô hình sẽ dự đoán nhóm khách hàng có khả năng cao nhất tham gia mua sắm trong tương lai. Nhằm giúp doanh nghiệp tối ưu hóa chiến lược tiếp thị và nâng cao hiệu quả trong việc tiếp cận để duy trì khách hàng. Với kết quả này, doanh nghiệp có thể đưa ra các chương trình khuyến mãi, ưu đãi nhằm gia tăng doanh thu, cải thiện trải nghiệm khách hàng, và tối ưu hóa chi phí bằng cách tập trung vào các nhóm khách hàng có giá trị tiềm năng cao nhất.

PHẦN NỘI DUNG

CHƯƠNG 1: GIỚI THIỆU

1.1. GIỚI THIỆU VỀ MACHINE LEARNING

Máy học (Machine Learning) là một nhánh của trí tuệ nhân tạo (AI) tập trung vào việc xây dựng các hệ thống có khả năng học hỏi từ dữ liệu và cải thiện hiệu suất theo thời gian mà không cần sự can thiệp trực tiếp của con người. Thay vì lập trình rõ ràng cụ thể như những cách giải quyết một vấn đề, các hệ thống sử dụng mô hình máy học để phân tích và "học" từ dữ liệu sau đó đưa ra quyết định hoặc dự đoán.

1.2. ỨNG DỤNG CỦA MACHINE LEARNING TRONG TIẾP THỊ VÀ TRONG THƯƠNG MẠI ĐIỆN TỬ

Máy học đang ngày càng trở thành công cụ quan trọng trong tiếp thị (marketing) và thương mại điện tử (e-commerce). Bằng cách sử dụng dữ liệu lớn (big data) và các thuật toán thông minh, máy học giúp cải thiện đáng kể các chiến lược tiếp thị và tối ưu hóa trải nghiệm mua sắm trực tuyến. Dưới đây là các ứng dụng chính của máy học trong lĩnh vực này:

1.2.1. Cá nhân hóa trải nghiệm khách hàng

Cá nhân hóa là một trong những ứng dụng quan trọng nhất của máy học trong tiếp thị và thương mại điện tử. Với sự trợ giúp của các thuật toán máy học, các doanh nghiệp có thể cung cấp trải nghiệm cá nhân hóa dựa trên hành vi và sở thích của từng khách hàng. Ví dụ:

- **Gợi ý sản phẩm**: Các hệ thống gợi ý mua sắm của người dùng và đề xuất các sản phẩm phù hợp dựa trên sở thích và lịch sử mua sắm của họ. Các trang thương mại điện tử như Amazon, Shopee, và Lazada thường sử dụng công nghệ này để tăng doanh số bán hàng.
- **Email Marketing cá nhân hóa**: Thay vì gửi cùng một nội dung đến tất cả khách hàng, máy học có thể giúp phân tích hành vi của từng người dùng để gửi thông tin quảng cáo đúng lúc, đúng nội dung mà họ quan tâm.

1.2.2. Dự đoán hành vi khách hàng

Máy học có thể phân tích lượng lớn dữ liệu từ hành vi duyệt web, lịch sử mua sắm, và các tương tác trên mạng xã hội để dự đoán các hành vi của khách hàng trong tương lai. Điều này bao gồm:

- **Dự đoán tỷ lệ mua hàng**: Xác định khả năng một khách hàng sẽ mua hàng sau khi thêm sản phẩm vào giỏ.
- **Phân tích khách hàng tiềm năng**: Dựa vào lịch sử tương tác và thói quen mua sắm, các doanh nghiệp có thể dự đoán ai là khách hàng tiềm năng và khi nào họ có khả năng thực hiện giao dịch.
- **Dự đoán xu hướng**: Nhận diện những thay đổi trong thị hiếu và xu hướng tiêu dùng để các doanh nghiệp có thể điều chỉnh chiến lược tiếp thị kịp thời.

1.2.3. Tối ưu hóa chiến lược quảng cáo

Máy học giúp tối ưu hóa việc đặt quảng cáo và phân bổ ngân sách quảng cáo một cách hiệu quả hơn. Bằng cách phân tích dữ liệu từ các chiến dịch quảng cáo trước đó, máy học có thể:

- **Tối ưu hóa đối tượng mục tiêu**: Xác định đối tượng khách hàng có tiềm năng cao nhất để tiếp cận và điều chỉnh chiến dịch tiếp thị cho phù hợp với họ.
- **Giảm chi phí quảng cáo**: Các thuật toán có thể được sử dụng để thử nghiệm các phiên bản quảng cáo khác nhau nhằm tối ưu hóa thông điệp và định dạng quảng cáo, từ đó giảm thiểu chi phí.

1.2.4. Chống gian lận

Trong thương mại điện tử, việc phát hiện và ngăn chặn gian lận là rất quan trọng. Máy học giúp phân tích các giao dịch và hành vi người dùng để phát hiện ra những hành vi bất thường. Các hệ thống này có thể học từ các mô hình giao dịch bình thường và cảnh báo khi phát hiện các hành vi đáng ngờ, như sử dụng thẻ tín dụng trái phép hoặc các hoạt động lừa đảo khác.

1.2.5. Quản lý chuỗi cung ứng và tồn kho

Máy học không chỉ giúp cải thiện trải nghiệm của khách hàng mà còn hỗ trợ các doanh nghiệp tối ưu hóa quản lý chuỗi cung ứng. Bằng cách phân tích các dữ liệu lịch sử về bán hàng, thời gian giao hàng và sự biến động của thị trường, các thuật toán máy học có thể:

- **Dự báo nhu cầu**: Dự đoán lượng hàng hóa cần thiết cho tương lai để tránh tình trạng thiếu hụt hoặc dư thừa tồn kho.
- **Tối ưu hóa quy trình vận chuyển**: Tìm ra những tuyến đường tốt nhất, giảm chi phí vận chuyển và tối ưu hóa thời gian giao hàng.

1.2.6. Chatbot và trợ lý ảo

Chatbot được hỗ trợ bởi máy học và xử lý ngôn ngữ tự nhiên (NLP) là một ứng dụng phổ biến trong thương mại điện tử. Các chatbot này có thể:

- **Hỗ trợ khách hàng tự động**: Trả lời các câu hỏi thường gặp, hỗ trợ khách hàng về đơn hàng, và hướng dẫn mua hàng.
- **Tăng cường dịch vụ khách hàng**: Phân tích ngôn ngữ và cảm xúc của khách hàng để đưa ra phản hồi phù hợp và cải thiện trải nghiệm tổng thể của họ.

1.3. GIỚI THIỆU VỀ DỮ LIỆU TRONG DỰ ÁN

1.3.1. Nguồn gốc của dữ liệu

Kaggle là một nền tảng trực tuyến nổi tiếng dành cho các nhà khoa học dữ liệu (data scientists) và những người đam mê học máy (machine learning). Được thành lập vào năm 2010 và hiện nay thuộc sở hữu của Google, Kaggle đã trở thành một điểm đến hàng đầu cho những ai muốn học hỏi, thực hành, và thi đấu trong các cuộc thi khoa học dữ liệu. Nền tảng này cung cấp các công cụ mạnh mẽ, bộ dữ liệu phong phú, và môi trường cộng tác để phát triển và chia sẻ các dự án học máy.

Nhóm em đã sử dụng Kaggle để lấy dữ liệu sử dụng cho việc phân tích và xây dựng mô hình.

Đường dẫn về dữ liệu được lấy trên Kaggle: Click tại đây

1.3.2. Mô tả dữ liệu

Dữ liệu này bao gồm thông tin chi tiết về khách hàng của một doanh nghiệp, cung cấp cái nhìn sâu sắc về nhiều khía cạnh khác nhau như thông tin cá nhân, hành vi mua sắm, cũng như mức độ tham gia của khách hàng trong các chiến dịch tiếp thị và quảng cáo.

Dữ liệu gồm có 29 cột và 2240 dòng.

Tên cột	Loại giá trị	Miền giá trị	Ý nghĩa
ID	Định lượng	0 - 11191	Mã số nhận dạng duy nhất cho mỗi khách hàng
Year_Birth	Định lượng	1893 - 1996	Năm sinh của khách hàng
Education	Phân loại	'Graduation' 'PhD''Master' 'Basic' '2n Cycle'	Trình độ học vấn của khách hàng
Marital_Status	Phân loại	['Single' 'Together' 'Married' 'Divorced' 'Widow' 'Alone' 'Absurd' 'YOLO']	Tình trạng hôn nhân của khách hàng
Income	Định lượng	1730.0 113734.0	Thu nhập hàng năm của khách hàng
Kidhome	Định lượng	0 - 2	Số trẻ em dưới 18 tuổi trong gia đình
Teenhome	Định lượng	0 - 2	Số thanh thiếu niên trong gia đình
Dt_Customer	Phân loại	01-01-2013 đến 31-12-2013	Ngày khách hàng trở thành khách hàng
Recency	Định lượng	0 - 99	Số ngày kể từ lần mua hàng cuối cùng của khách hàng
MntWines	Định lượng	0 - 1493	Số tiền đã chi tiêu cho rượu vang trong 2 năm gần nhất

MntFruits	Định lượng	0 - 199	Số tiền đã chi tiêu cho trái cây trong 2 năm gần nhất
MntMeatProducts	Định lượng	0 - 1725	Số tiền đã chi tiêu cho sản phẩm thịt trong 2 năm gần nhất
MntFishProducts	Định lượng	0 - 259	Số tiền đã chi tiêu cho sản phẩm cá trong 2 năm gần nhất
MntSweetProducts	Định lượng	0 - 263	Số tiền đã chi tiêu cho sản phẩm ngọt trong 2 năm gần nhất
MntGoldProds	Định lượng	0 - 362	Số tiền đã chi tiêu cho sản phẩm vàng trong 2 năm gần nhất
NumDealsPurchases	Định lượng	0 - 15	Số lần mua hàng có khuyến mãi trong 2 năm gần nhất
NumWebPurchases	Định lượng	0 - 27	Số lần mua hàng qua website trong 2 năm gần nhất
NumCatalogPurchases	Định lượng	0 - 28	Số lần mua hàng qua catalog trong 2 năm gần nhất
NumStorePurchases	Định lượng	0 - 13	Số lần mua hàng trực tiếp tại cửa hàng trong 2 năm gần nhất
NumWebVisitsMonth	Định lượng	0 - 20	Số lần truy cập website trong tháng gần nhất
AcceptedCmp3	Phân loại	0, 1	Chấp nhận chiến dịch thứ 3 (1: Có, 0: Không)

AcceptedCmp4	Phân loại	0, 1	Chấp nhận chiến dịch thứ 4 (1: Có, 0: Không)
AcceptedCmp5	Phân loại	0, 1	Chấp nhận chiến dịch thứ 5 (1: Có, 0: Không)
AcceptedCmp1	Phân loại	0, 1	Chấp nhận chiến dịch thứ 1 (1: Có, 0: Không)
AcceptedCmp2	Phân loại	0, 1	Chấp nhận chiến dịch thứ 2 (1: Có, 0: Không)
Complain	Phân loại	0, 1	Khiếu nại (1: Có, 0: Không)
Z_CostContact	Định lượng	3	Chi phí liên hệ marketing
Z_Revenue	Định lượng	11	Doanh thu
Response	Phân loại	0, 1	Phản hồi tích cực với chiến dịch (1: Có, 0: Không)

1.3.3. Những vấn đề trong dữ liệu

Thiếu dữ liệu:

Cột "Income" (Thu nhập) có 24 giá trị bị thiếu. Đây là một số lượng nhỏ so với tổng 2240 bản ghi, nhưng vẫn cần được xử lý vì cột này có thể quan trọng trong các phân tích hoặc mô hình học máy.

Giá trị ngoại lai:

Sử dụng phương pháp IQR (Interquartile Range), đã phát hiện ngoại lai trong các cột liên quan đến chi tiêu:

- **MntWines:** 35 giá trị ngoại lai.
- **MntFruits:** 227 giá trị ngoại lai.

- **MntMeatProducts:** 175 giá trị ngoại lai.
- **MntFishProducts:** 223 giá trị ngoại lai.
- **MntSweetProducts:** 248 giá trị ngoại lai.
- **MntGoldProds:** 207 giá trị ngoại lai.

Dữ liệu mất cân bằng:

Các cột nhị phân liên quan đến việc khách hàng tham gia các chiến dịch tiếp thị (**AcceptedCmp1** đến **AcceptedCmp5**) và phản hồi của khách hàng (**Response**) có sự mất cân bằng. Hơn **93% đến 98%** khách hàng **không** tham gia các chiến dịch tiếp thị và hơn **85%** khách hàng **không** phản hồi lại các chiến dịch.

Dữ liệu mất cân bằng như thế này có thể gây khó khăn trong việc xây dựng các mô hình phân loại vì các mô hình có thể thiên lệch về phía các lớp chiếm đa số.

CHƯƠNG 2: PHƯƠNG PHÁP VÀ CÔNG CỤ SỬ DỤNG

2.1. XỬ LÝ DỮ LIỆU

- Xử lý dữ liệu thiếu, giá trị ngoại lai (outliers):
+ Loại bỏ các dòng hoặc cột có quá nhiều giá trị thiếu.
+ Sử dụng biểu đồ hộp (boxplot) để nhận diện ngoại lai.
- Gộp các cột có sự tương đồng lại với nhau để giảm chiều dữ liệu.
    + Số trẻ
    + Số tiền chi tiêu cho các loại sản phẩm khác nhau
    + Số lượng giao dịch qua các kênh khác nhau
    + Chỉ số cho biết khách hàng có nhập nhận các chiến dịch
- Mã hóa các biến định tính (categorical variables): Các biến định tính cần được mã hóa để sử dụng trong mô hình học máy. Các phương pháp mã hóa phổ biến bao gồm:
+ Label Encoding: Chuyển các nhãn định tính thành các giá trị số nguyên (0, 1, 2,...).
+ One-Hot Encoding: Tạo các cột nhị phân cho mỗi giá trị trong biến định tính để tránh ảnh hưởng thứ tự trong mô hình học máy.

2.2. CÁC MÔ HÌNH LỰA CHỌN

2.2.1. Giải thích về mô hình

- **K-Means:** là thuật toán phân cụm, giúp chia khách hàng thành các nhóm dựa trên sự tương đồng. Thuật toán chọn số lượng cụm trước (k) và phân chia các điểm dữ liệu vào các cụm khác nhau cho đến khi ổn định.
- **Logistic Regression:** là mô hình phân loại, dùng để dự đoán khả năng một khách hàng thuộc vào nhóm tiềm năng mua sắm hay không. Nó tính toán xác suất và phân loại dựa trên ngưỡng (thường là 0.5).
- **Random Forest:** là một tập hợp nhiều cây quyết định. Mỗi cây đưa ra dự đoán và kết quả cuối cùng là sự kết hợp của các dự đoán này. Mô hình này rất mạnh mẽ, tránh được lỗi quá khớp và phù hợp cho các bài toán phức tạp.

2.2.2. Lý do chọn mô hình

- **K-Means** được chọn vì nó giúp phân nhóm khách hàng dựa trên những đặc điểm chung, hỗ trợ doanh nghiệp xác định các nhóm khách hàng tiềm năng.
- **Logistic Regression** phù hợp với bài toán phân loại nhị phân, trong đó ta dự đoán khách hàng có khả năng mua hàng hay không. Mô hình này đơn giản và dễ triển khai.
- **Random Forest** là lựa chọn tốt vì tính chính xác cao, khả năng xử lý các đặc điểm không tuyến tính, và đặc biệt hiệu quả trong việc tránh quá khớp dữ liệu.

2.2.3. Huấn luyện mô hình

- **K-Means**:
  Đầu tiên, lựa chọn số lượng cụm (**k**) thích hợp bằng cách sử dụng các phương pháp như **Elbow Method** hoặc **Silhouette Score** để xác định điểm tối ưu, nơi mà việc tăng số lượng cụm không còn cải thiện chất lượng phân cụm đáng kể. Sau đó, mô hình K-Means được huấn luyện bằng cách phân chia các khách hàng vào những cụm khác nhau dựa trên khoảng cách từ điểm dữ liệu đến các tâm cụm (centroids). Quá trình huấn luyện tiếp tục với việc cập nhật tâm cụm cho đến khi đạt sự ổn định, giúp phân nhóm khách hàng tiềm năng dựa trên hành vi và đặc điểm tương tự nhau.
- **Logistic Regression**:
  Đối với mô hình Logistic Regression, sau khi chuẩn bị và tiền xử lý dữ liệu, tập huấn luyện sẽ được sử dụng để huấn luyện mô hình. Mô hình Logistic Regression tìm kiếm mối quan hệ giữa các đặc điểm của khách hàng (biến độc lập) và khả năng mua sắm (biến phụ thuộc). Bằng cách áp dụng hàm **sigmoid**, mô hình chuyển các giá trị đầu vào thành xác suất và từ đó phân loại khách hàng có khả năng mua hàng hay không. Việc huấn luyện này sẽ giúp mô hình tối ưu hóa các tham số để đạt được dự đoán tốt nhất.
- **Random Forest**:
  Mô hình Random Forest được huấn luyện bằng cách tạo ra nhiều cây quyết định từ các mẫu dữ liệu ngẫu nhiên. Mỗi cây trong "rừng" sẽ thực hiện dự đoán riêng dựa trên đặc điểm của khách hàng, sau đó kết hợp các kết quả từ tất cả các cây để đưa ra dự đoán cuối cùng. Mô hình này không chỉ tăng độ chính xác mà còn giúp giảm thiểu hiện tượng **overfitting** (quá khớp), nhờ vào khả năng trung bình hóa kết quả từ nhiều cây. Trong quá trình huấn luyện, Random Forest liên tục được cải thiện để đưa ra các dự đoán phân loại khách hàng tiềm năng chính xác hơn, giúp xác định nhóm khách hàng có nhiều khả năng mua hàng nhất.
- **Mô hình phân loại kết hợp phân cụm (LogisticRegression - K-Means)**:
  kết hợp cả hai kỹ thuật phân cụm và phân loại để cải thiện khả năng dự đoán và phân tích dữ liệu.

CHƯƠNG 3: ĐÁNH GIÁ KẾT QUẢ

3.1. CẢI THIỆN CHẤT LƯỢNG DỮ LIỆU VÀ ĐẶC TRƯNG

3.1.1. Lựa chọn đặc trưng với SelectKBest

- Phương pháp SelectKBest giúp xác định các đặc trưng có mối tương quan mạnh nhất với biến mục tiêu, từ đó ưu tiên các đặc trưng quan trọng.
- Giảm số chiều dữ liệu bằng cách loại bỏ các đặc trưng không liên quan hoặc nhiễu.
- Quá trình này cải thiện hiệu suất của mô hình và tăng tốc độ huấn luyện bằng cách tập trung vào các đặc trưng có tác động lớn nhất.
- Trong phương pháp này, f_classif (ANOVA F-test) được sử dụng để đánh giá và lựa chọn các đặc trưng dựa trên ý nghĩa thống kê của chúng.

3.1.2. Sinh dữ liệu với Conditional Generative Adversarial Network (CGAN)

1. Discriminator (D Loss) ~1.2:

- Discriminator hoạt động ổn định, duy trì khả năng phân biệt dữ liệu thật và giả một cách hợp lý.
- Mức D Loss quanh 1.2 là dấu hiệu cho thấy Discriminator đủ mạnh nhưng không quá áp đảo, đảm bảo sự cân bằng với Generator.

2. Generator (G Loss) ~1.1:

- Generator đã học được cách sinh dữ liệu giả với chất lượng cao và khó phân biệt với dữ liệu thật.
- Mức G Loss > 1.0 cho thấy Generator đang hoạt động tốt, đánh lừa Discriminator hiệu quả.

3. Tổng quan mô hình

- Trạng thái cân bằng: Cả Discriminator và Generator đều đạt sự cân bằng tốt, không thành phần nào áp đảo thành phần còn lại, giúp mô hình học hiệu quả.
- Chất lượng dữ liệu giả sinh ra: Với D Loss và G Loss như hiện tại, dữ liệu giả được kỳ vọng có độ chân thực cao, phân phối sát với dữ liệu thật.
- Hội tụ: Các giá trị D Loss và G Loss ổn định qua các epoch, cho thấy mô hình CGAN đã hội tụ.

3.2. PHÁT TRIỂN MÔ HÌNH

3.2.1. Cải thiện hiệu suất phân cụm với K-means

Chuẩn hóa dữ liệu: Sử dụng StandardScaler để đảm bảo các đặc trưng có cùng phạm vi giá trị, giảm ảnh hưởng của thang đo khác biệt.

Xác định số cụm tối ưu:

- Elbow Method: Tìm điểm "khuỷu tay" từ biểu đồ SSD.
- Silhouette Score: Đánh giá mức độ tách biệt và gắn kết giữa các cụm.

Tinh chỉnh thuật toán:

- Sử dụng k-means++ để khởi tạo cụm tốt hơn.
- Tăng số lần chạy lại (n_init) để đảm bảo kết quả ổn định.

3.2.2. So sánh các mô hình với LazyPredict

- Đánh giá nhanh hiệu suất của nhiều mô hình học máy
- Logistic Regression cho kết quả tốt nhất về độ chính xác và thời gian huấn luyện
- Xác định được các mô hình tiềm năng cho việc phát triển tiếp theo

3.2.3. Tối ưu hóa Logistic Regression

- Cross-validation để đảm bảo độ ổn định của mô hình
- Tìm kiếm siêu tham số tối ưu thông qua GridSearchCV

3.2.4. Kết hợp K-means với mô hình phân loại (Logistic Regression)

- Phân cụm khách hàng thành các nhóm có đặc điểm tương đồng
- Sử dụng thông tin phân cụm làm đặc trưng bổ sung cho mô hình phân loại
- Tăng khả năng giải thích của mô hình về hành vi khách hàng

3.3. KẾT QUẢ ĐẠT ĐƯỢC

3.3.1. Mô hình phân cụm với K-means

Việc cải thiện điểm Silhouette Score từ 0.1988 lên 0.3485 cho thấy rõ ràng rằng:

- Chọn đặc trưng phù hợp đã giúp tập trung vào các yếu tố quan trọng, loại bỏ nhiễu, và tăng cường khả năng tách biệt giữa các cụm.
- Sinh dữ liệu mới bằng CGAN đã bổ sung sự đa dạng vào tập dữ liệu, cung cấp thêm thông tin hữu ích, giúp thuật toán K-means xác định ranh giới giữa các cụm tốt hơn.

3.3.2. So sánh với kết quả khi chưa cải thiện dữ liệu

- Độ chính xác tăng từ 88% lên 99%
- Precision cải thiện từ 78% lên 99%
- Cross-validation score ổn định hơn với độ lệch chuẩn thấp hơn

3.3.3. Ý nghĩa thực tiễn

- Dự đoán chính xác hơn các khách hàng tiềm năng
- Giảm chi phí marketing không hiệu quả
- Cung cấp insight về các nhóm khách hàng khác nhau

CHƯƠNG 4: KẾT LUẬN

4.1. TỔNG KẾT NGHIÊN CỨU

Cải thiện chất lượng dự đoán: Việc kết hợp K-means clustering với Logistic Regression giúp cải thiện đáng kể độ chính xác trong việc phân loại khách hàng tiềm năng. Phương pháp phân cụm K-means giúp nhóm khách hàng theo các đặc điểm hành vi tương tự, từ đó tạo ra các nhóm phân loại rõ ràng hơn. Điều này cho phép mô hình phân loại các khách hàng có xu hướng phản hồi cao đối với các chiến dịch tiếp thị, giúp doanh nghiệp tối ưu hóa chiến lược marketing bằng cách nhắm mục tiêu chính xác hơn vào các nhóm khách hàng có khả năng cao tham gia vào các chiến dịch.

Phân tích hành vi khách hàng: Qua phân cụm khách hàng theo các yếu tố như độ tuổi, thu nhập và chi tiêu, nghiên cứu đã phân chia khách hàng thành các nhóm hành vi khác nhau. Các nhóm này có những đặc điểm và nhu cầu riêng biệt. Ví dụ, nhóm trung niên có xu hướng chi tiêu nhiều cho các sản phẩm gia đình, trong khi nhóm trẻ tuổi lại tập trung vào các sản phẩm cá nhân. Nhờ vào những phân tích này, các chiến lược marketing có thể được thiết kế phù hợp với từng nhóm, tập trung vào các sản phẩm và dịch vụ đáp ứng đúng nhu cầu và mong muốn của khách hàng, từ đó tăng cường sự hài lòng và khả năng mua sắm của họ.

Ứng dụng vào chiến lược marketing: Kết quả từ nghiên cứu giúp doanh nghiệp tối ưu hóa chiến lược tiếp thị bằng cách xác định chính xác nhóm khách hàng tiềm năng và tạo ra các chiến lược quảng cáo hiệu quả hơn. Bằng cách tập trung vào các nhóm có khả năng phản hồi cao, doanh nghiệp có thể tiết kiệm chi phí quảng cáo, nâng cao tỷ lệ chuyển đổi và tối ưu hóa nguồn lực cho các chiến dịch marketing, từ đó gia tăng doanh thu và cải thiện mối quan hệ với khách hàng.

4.2. HƯỚNG PHÁT TRIỂN

Thử nghiệm các kỹ thuật sinh dữ liệu khác: Thực hiện các phương pháp sinh dữ liệu như GANs hoặc SMOTE để tạo ra dữ liệu bổ sung, giúp tăng cường độ chính xác của mô hình, đặc biệt là trong các trường hợp dữ liệu không cân bằng.

Phát triển mô hình ensemble: Kết hợp nhiều mô hình học máy (như Random Forest và XGBoost) để cải thiện độ chính xác dự đoán và giảm thiểu sai số trong quá trình phân loại khách hàng tiềm năng.

Tích hợp thêm dữ liệu từ nhiều nguồn: Sử dụng dữ liệu từ các nguồn khác nhau như mạng xã hội và giao dịch trực tuyến để làm phong phú thêm các dự đoán, giúp mô hình hiểu rõ hơn về hành vi khách hàng.

Xây dựng hệ thống dự đoán real-time: Phát triển hệ thống có khả năng dự đoán và cập nhật ngay lập tức khi có dữ liệu mới, giúp tối ưu hóa các chiến dịch tiếp thị và phản hồi nhanh chóng với thay đổi của khách hàng.

Tối ưu hóa siêu tham số mô hình: Áp dụng các kỹ thuật tối ưu hóa siêu tham số như Grid Search và Bayesian Optimization để cải thiện hiệu suất của mô hình và đạt được kết quả dự đoán chính xác hơn.