**VIETNAM NATIONAL UNIVERSITY - VNUHCM**

**HO CHI MINH CITY UNIVERSITY OF INFORMATION TECHONOLOGY**

A blue logo with a black background

Description automatically generated with low confidence

**ĐỒ ÁN MÔN HỌC**

**MÔN HỌC: DỮ LIỆU LỚN**

**LỚP: IS405.O11.HTCL**

**ĐỀ TÀI: ỨNG DỤNG PHÂN TÍCH DỮ LIỆU LỚN TRONG PHÂN KHÚC KHÁCH HÀNG SỬ DỤNG THUẬT TOÁN KMEANS**

**GIẢNG VIÊN:**

**ThS. Nguyễn Hồ Duy Tri**

**SINH VIÊN:**

|  |  |
| --- | --- |
| **Nguyễn Quốc Khánh** | **20521452** |
| **Bùi Thị Thanh Ngân** | **20521643** |
| **Nguyễn Thị Phương Nhi** | **19520814** |
| **Phạm Lê Trưởng** | **20522090** |

**Hồ Chí Minh, 12/2023**

**MỤC LỤC**

[**LỜI MỞ ĐẦU 3**](#_Toc154034782)

[**LỜI CẢM ƠN 4**](#_Toc154034783)

[**CHƯƠNG 1: GIỚI THIỆU 5**](#_Toc154034784)

[**1.1. PHÁT BIỂU BÀI TOÁN 5**](#_Toc154034785)

[**1.2. TỔNG QUAN DỮ LIỆU 6**](#_Toc154034786)

[**1.2.1. Mô tả về dữ liệu 6**](#_Toc154034787)

[**1.2.2. Thuộc tính của kho dữ liệu 6**](#_Toc154034788)

[**CHƯƠNG 2: PHẦN MỀM VÀ CÔNG CỤ 7**](#_Toc154034789)

[**2.1. Framework 7**](#_Toc154034791)

[**2.1.1. Apache Hadoop 7**](#_Toc154034792)

[**2.1.2. Apache Spark 8**](#_Toc154034793)

[**2.2. Môi trường 8**](#_Toc154034794)

[**2.3. Phần mềm 8**](#_Toc154034795)

[**CHƯƠNG 3: XỬ LÝ DỮ LIỆU LỚN 9**](#_Toc154034796)

[**3.1. Tiền xử lý dữ liệu 9**](#_Toc154034798)

[**CHƯƠNG 4: GIẢI THUẬT VÀ SONG SONG HÓA GIẢI THUẬT 21**](#_Toc154034799)

[**4.1. Thuật toán Kmeans 21**](#_Toc154034801)

[**4.1.1. Giới thiệu về thuật toán Kmeans 21**](#_Toc154034802)

[**4.1.2. Ý tưởng thuật toán Kmeans 21**](#_Toc154034803)

[**4.1.3. Điều kiện hội tụ 22**](#_Toc154034804)

[**4.1.4. Xác định điểm trung tâm của cluster 22**](#_Toc154034805)

[**4.1.5. Phép đo khoảng cách 23**](#_Toc154034806)

[**4.2. Một số yếu tố ảnh hưởng đến thuật toán Kmeans 23**](#_Toc154034807)

[**4.2.1. Ảnh hưởng của outlier 23**](#_Toc154034808)

[**4.2.2. Ảnh hưởng của việc khởi tạo trung tâm 24**](#_Toc154034809)

[**4.3. Mô hình song song hóa 25**](#_Toc154034810)

[**4.4. Xây dựng mô hình Kmeans với song song hóa giải thuật 25**](#_Toc154034811)

[**4.4.1. Xây dựng mô hình Kmeans 25**](#_Toc154034812)

[**4.4.2. Đánh giá mô hình 26**](#_Toc154034813)

[**CHƯƠNG 5: KẾT QUẢ VÀ HƯỚNG PHÁT TRIỂN 28**](#_Toc154034814)

[**5.1. KẾT QUẢ VÀ KẾT LUẬN 28**](#_Toc154034816)

[**5.1.1. RFM Model (Recency - Frequency - Monetary Model) 28**](#_Toc154034817)

[**5.1.2. Kmeans Clustering 33**](#_Toc154034818)

[**5.2. HƯỚNG PHÁT TRIỂN 37**](#_Toc154034819)

[**CHƯƠNG 6: THAM KHẢO 39**](#_Toc154034820)

# 

# LỜI MỞ ĐẦU

Cùng với xu thế toàn cầu hóa và hội nhập quốc tế đang phát triển mạnh mẽ, ngày càng có nhiều dữ liệu được thu thập và lưu trữ dưới nhiều dạng khác nhau, gần hàng chục triệu tấm ảnh được lưu trữ mỗi ngày và hàng ngàn terabyte được sử dụng mỗi giây. Do đó, việc phân tích dữ liệu và đưa ra những dự đoán cần thiết là hết sức quan trọng và thiết yếu đối với xã hội ngày nay, từ đó giúp doanh nghiệp cũng như nhiều hoạt động quan trọng khác biết được hướng đi và nắm rõ sứ mệnh của mình.

Vì lý do đó nhóm em quyết định tìm hiểu về phân tích dữ liệu lớn và chọn kho dữ liệu về bán lẻ, dựa trên hoá đơn từ đó có thể dự đoán những hành vi của khách hàng. Để bước đầu tìm hiểu về dữ liệu và phân tích dữ liệu dựa trên Spark và hệ thống tính toán song song.

# LỜI CẢM ƠN

Trước khi đi vào nội dung phần báo cáo đồ án, đầu tiên xin cho phép nhóm của chúng em gửi lời cảm ơn sâu sắc đến thầy Nguyễn Hồ Duy Tri - giảng viên hướng dẫn môn học Dữ liệu lớn (IS405.O11.HTCL) đã cung cấp cho chúng em những kiến thức bổ ích và sự trợ giúp cần thiết trong suốt khoảng thời gian thực hiện đồ án, cũng như khoa Hệ thống Thông tin, Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh đã tạo những điều kiện tốt nhất giúp cho chúng em có được cơ hội để thực hiện hoá đề tài và hoàn thiện đồ án này. Cuối cùng, chúng em xin kính chúc Quý Thầy/Cô của khoa Hệ thống Thông tin cũng như thầy Nguyễn Hồ Duy Tri sức khoẻ dồi dào và thành công trên lĩnh vực của mình để sẵn sàng tiếp tục trên con đường truyền đạt kiến thức, truyền lửa và nhiệt huyết cho thế hệ mai sau. Xin trân trọng cảm ơn Quý Thầy/Cô rất nhiều

# CHƯƠNG 1: GIỚI THIỆU

## PHÁT BIỂU BÀI TOÁN

Theo nguyên lý pareto 20% khách hàng sẽ mang lại 80% doanh số. Do đó doanh nghiệp cần xác định được những khách hàng quan trọng nhất để chăm sóc đặc biệt. Những tập khách hàng này được gọi là VIP, Priority hoặc khách hàng cao cấp, tùy từng doanh nghiệp có cách gọi khác nhau. Việc phân chia khách hàng thành các nhóm khác nhau dựa trên nhu cầu mua sắm sẽ giúp doanh nghiệp kinh doanh hiệu quả hơn, marketing target đến đúng tập khách hàng hơn, và khách hàng được phục vụ tốt hơn.

Doanh nghiệp sẽ chia nhỏ tệp khách hàng ra thành những nhóm đối tượng khác nhau để có thể áp dụng những chiến lược kinh doanh cụ thể cho từng nhóm đối tượng. Điều này giúp cho khách hàng được tiếp cận với các sản phẩm thật sự phù hợp với bản thân họ. Sự phù hợp đó sẽ kéo doanh số tăng lên. Vấn đề đặt ra là làm sao có thể chia nhỏ tệp khách hàng đó ra khi mà số lượng hóa đơn là rất lớn và không thể phân tích từng vị khách. Và mục tiêu của các bài toán là từ tập dữ liệu khổng lồ đó. Làm sao biết được mỗi tệp khách hàng có những đặc trưng nào? Đó chính là vấn đề mà bài toán cần giải quyết.

Phân khúc khách hàng được nhóm thực hiện thông qua hai phương pháp khác nhau:

* RFM Model: Tính trung bình rank các biến: tính trung bình ranks của các biến ta thu được giá trị rank tổng hợp ứng với mỗi khách hàng. Ta có thể giữ nguyên mỗi rank tổng hợp là một nhóm hoặc gộp nhiều rank thành một nhóm theo khoảng giá trị như bên dưới:
* Khách hàng VIP: rank từ 8-10.
* Khách hàng đại chúng: rank từ 5-7.
* Khách hàng thứ cấp: rank < 5.
* Sử dụng thuật toán K-mean clustering để phân cụm các khách hàng dựa vào input là 3 biến giá trị ranks của Recency, Frequency, Monetary. Theo cách này ta sẽ tự động phân cụm toàn bộ khách hàng về 3 nhóm mà không cần phải tạo khoảng rank. Tuy nhiên hạn chế là các clustering không có một định nghĩa rõ ràng (VIP, đại chúng, thứ cấp).

Nhóm sẽ phân tích các giao dịch trực tuyến ở UK và các quốc gia khác đang mua hàng từ cửa hàng bán đồ quà tặng. Sử dụng mô hình Machine Learning, từ đó phân tích đánh giá hành vi mua hàng của khách hàng thông qua:

* Số tiền khách hàng đã chi cho cửa hàng
* Số hoá đơn của 1 khách hàng đã có
* Số lượng hàng hoá mà khách hàng đã mua

Trong đồ án này, nhóm em sử dụng dataset với các thông số như trên để đưa ra những khách hàng tiềm năng cho cửa hàng

## TỔNG QUAN DỮ LIỆU

### Mô tả về dữ liệu

Đây là bộ dữ liệu xuyên quốc gia chứa tất cả các giao dịch xảy ra từ ngày 12/01/2010 đến ngày 12/09/2011 của một cửa hàng bán lẻ trực tuyến và có trụ sở tại Vương quốc Anh. Công ty chủ yếu bán quà tặng độc đáo cho mọi dịp, lễ. Nhiều khách hàng của công ty là những người bán buôn.

Kho dữ liệu gồm 1 bảng, với 541909 dòng và 8 thuộc tính

Link dataset: [E-Commerce Data (kaggle.com)](https://www.kaggle.com/datasets/carrie1/ecommerce-data/data)

### Thuộc tính của kho dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ý nghĩa** |
| 1 | InvoiceNo | String | Số hoá đơn |
| 2 | StockCode | String | Mã hàng hoá |
| 3 | Description | String | Mô tả hàng hoá |
| 4 | Quantity | Integer | Số lượng hàng hoá đã mua |
| 5 | InvoiceDate | String | Ngày in hoá đơn |
| 6 | UnitPrice | Float | Giá |
| 7 | CustomerID | Integer | ID Khách hàng |
| 8 | Country | String | Quốc gia |

# CHƯƠNG 2: PHẦN MỀM VÀ CÔNG CỤ



## Framework

### Apache Hadoop

Hadoop là một Big Data framework, cung cấp khả năng phân phối lưu trữ dữ liệu, xử lý song song và xử lý dữ liệu với khối lượng lớn hơn, tốc độ cao hơn, đa dạng, giá trị và tính xác thực. HDFS, MapReduce và YARN là ba thành phần chính trong Hadoop

* Hadoop HDFS sử dụng các nút tên và nút dữ liệu để lưu trữ dữ liệu mở rộng
* MapReduce quản lý các nút này để xử lý
* YARN hoạt động như một hệ điều hành cho Hadoop trong việc quản lý tài nguyên cụm

### Apache Spark

Apache Spark là một framework xử lý phân tán dữ liệu mã nguồn mở được sử dụng trên quy mô lớn và cung cấp các API cho Java, Scala, Python và R. Bằng cách sử dụng bộ nhớ đệm, quyền truy vấn và các mã để thực hiện nhiều công việc như phân tích nhanh dữ liệu, xử lý quy mô, truy vấn tương tác hay xử lý đồ hoạ

Spark không sử dụng hệ thống lưu trữ riêng mà chạy phân tích trên các hệ thống như HDFS hoặc bên thứ ba như Amazon Redshift, Amazon S3, Couchbase, Cassandra. Spark trên Hadoop sử dụng YARN để chia sẻ một cụm và tập dữ liệu chung đảm bảo mức độ dịch vụ và phản hồi nhất quán.

## Môi trường

Ubuntu là một hệ điều hành máy tính mã nguồn mở, được xây dựng trên nền tảng Linux. Nó được phát triển bởi Canonical Ltd. Ubuntu được thiết kế để dễ sử dụng và là một trong những phiên bản Linux phổ biến nhất trên thế giới

Đặc điểm nổi bật của Ubuntu bao gồm giao diện người dùng thân thiện, khả năng tương thích tốt với nhiều phần cứng, quản lý gói phần mềm dựa trên hệ thống quản lý gói Debian, và cung cấp môi trường tích hợp để phát triển ứng dụng. Ngoài ra, Ubuntu cũng được cập nhật đều đặn với các phiên bản mới và bảo mật, giúp người dùng duy trì một hệ thống an toàn và hiệu suất tốt.

## Phần mềm

Jupyter Notebook là một ứng dụng web mở rộng cho việc tạo và chia sẻ tài liệu tương tác chứa mã nguồn, văn bản có định dạng, hình ảnh, công thức toán học và các yếu tố tương tác khác. Đặc điểm nổi bật của Jupyter Notebook là khả năng kết hợp mã nguồn thực thi và nội dung mô tả trong cùng một tài liệu.

Người sử dụng có thể viết và thực thi mã từng phần (cell) một, giúp họ kiểm tra từng đoạn mã một cách linh hoạt. Kết quả của mã thực thi cũng có thể được hiển thị ngay trong tài liệu, kèm theo hình ảnh, đồ thị và các phương tiện trực quan khác.

Jupyter được đặt tên dựa trên ba ngôn ngữ lập trình chính mà nó hỗ trợ: Julia, Python và R. Tuy nhiên, nó cũng hỗ trợ nhiều ngôn ngữ khác thông qua các kernel riêng biệt được cài đặt.

# CHƯƠNG 3: XỬ LÝ DỮ LIỆU LỚN



## Tiền xử lý dữ liệu

Bước 1: Thêm 1 số thư viện cần thiết

A black and purple text

Description automatically generated

Bước 2:Build SparkSession và định nghĩa lại Schema cho Dataframe

A screenshot of a computer program

Description automatically generated

Bước 3: Tiến hành đọc dữ liệu

A screenshot of a computer

Description automatically generated

Bước 4: Xem các giá trị của dữ liệu (min,max,count)

A close-up of a computer screen

Description automatically generated

Bước 5: Xem và xóa các giá trị rỗng

A screenshot of a computer

Description automatically generated

Bước 6: Loại bỏ những data row có Quantity và UnitPrice có giá trị âm, loại bỏ data trùng lặp.

A screenshot of a computer code

Description automatically generated

Bước 7 : Xem kích thước dữ liệu

A close-up of a computer code

Description automatically generated

Bước 8: Tạo cột Recency bằng cách lấy ra ngày mua hàng gần nhất – InvoiceDate

A screenshot of a computer

Description automatically generated

Kết quả sau khi tạo cột Recency

A table of numbers and numbers

Description automatically generated with medium confidence

Xem kết quả count, min, max của cột rerency

A screenshot of a computer

Description automatically generated

Vẽ biểu đồ cho Recency

A screen shot of a graph

Description automatically generated

Bước 9: Tạo cột Frequency

Tạo df\_freq Frequency từ 2 cột CustomerID và InvoiceDate với mỗi khách hàng tiến hành countDistinct cột InvoiceDate để biết được tần suất mua hàng của khách.

A close-up of a computer screen

Description automatically generated

Vẽ biểu đồ cho Frequency

A screen shot of a graph

Description automatically generated

Bước 10: Tiến hành tạo cột Monetary bằng cách lấy Quantity x UnitPrice groupBy theo từng khách hàng

A screenshot of a computer code

Description automatically generated

Vẽ biểu đồ cho Monetary mức chi tiêu của khách hàng

A graph of a distribution of total amount

Description automatically generated

Bước 10: Tiến hành tạo TempView cho 3 dataframe và kết 3 bảng dữ liệu lại

A close-up of a white background

Description automatically generated

A screenshot of a computer code

Description automatically generated

Sau khi kết và loại bỏ một số cột không cần thiết được dữ liệu gồm 5 cột CustomerID,Country,Recency,Frequency và Monetary

A screenshot of a computer

Description automatically generated

Bước 11: Tiến hành Encode cho cột Country

A screenshot of a computer

Description automatically generated

Bước 12: Lấy các input\_feature gồm 4 cột Country\_Encode,Recency,Frequency, Monetary

A screenshot of a computer

Description automatically generated

Bước 12: Tiến hành tranform dữ lệu sang dạng vector và scaling dữ liệu

A screenshot of a computer program

Description automatically generated

A computer code with black text

Description automatically generated

Bước 13: Cuối cùng giảm chiều của dữ liệu bằng cách sử dụng PCA và tiến hành select cột feature\_scaled\_pca vào df để chuẩn bị đưa vào modelA screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

# CHƯƠNG 4: GIẢI THUẬT VÀ SONG SONG HÓA GIẢI THUẬT



## Thuật toán Kmeans

### Giới thiệu về thuật toán Kmeans

K-means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát (tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau.

Thuật toán phân cụm k-means thường được sử dụng trong các ứng dụng cỗ máy tìm kiếm, phân đoạn khách hàng, thống kê dữ liệu,…

Trong thuật toán k-means mỗi cụm dữ liệu được đặc trưng bởi một tâm (centroid). Tâm là điểm đại diện nhất cho một cụm và có giá trị bằng trung bình của toàn bộ các quan sát nằm trong cụm. Chúng ta sẽ dựa vào khoảng cách từ mỗi quan sát tới các tâm để xác định nhãn cho chúng trùng thuộc về tâm gần nhất. Ban đầu thuật toán sẽ khởi tạo ngẫu nhiên một số lượng xác định trước tâm cụm. Sau đó tiến hành xác định nhãn cho từng điểm dữ liệu và tiếp tục cập nhật lại tâm cụm. Thuật toán sẽ dừng cho tới khi toàn bộ các điểm dữ liệu được phân về đúng cụm hoặc số lượt cập nhật tâm chạm ngưỡng.

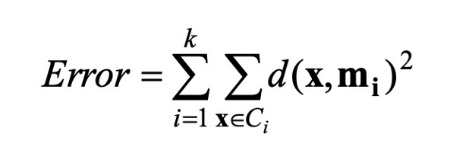
### Ý tưởng thuật toán Kmeans

1. Khởi tạo K điểm dữ liệu trong bộ dữ liệu và tạm thời coi nó là tâm của các cụm dữ liệu của chúng ta.
2. Với mỗi điểm dữ liệu trong bộ dữ liệu, tâm cụm của nó sẽ được xác định là 1 trong K tâm cụm gần nó nhất.
3. Sau khi tất cả các điểm dữ liệu đã có tâm, tính toán lại vị trí của tâm cụm để đảm bảo tâm của cụm nằm ở chính giữa cụm.
4. Bước 2 và bước 3 sẽ được lặp đi lặp lại cho tới khi vị trí của tâm cụm không thay đổi hoặc tâm của tất cả các điểm dữ liệu không thay đổi.

### Điều kiện hội tụ

Ta sẽ xác định điều kiện dừng thuật toán theo một số cách như sau:

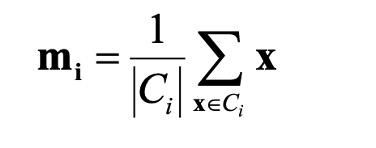
* Tại 1 vòng lặp: có ít các điểm dữ liệu được gán sang cluster khác hoặc
* Điểm trung tâm (centroid) không thay đổi nhiều hoặc
* Giá trị hàm mất mát không thay đổi nhiều:



Trong đó  là cluster thứ i,  là điểm trung tâm của cluster  tương ứng.

Nhìn chung về điều kiện hội tụ có thể thấy mối liên hệ giữa các điều kiện là gần tương đồng như nhau. Khi có ít điểm dữ liệu được gán sang cluster khác có thể khiến điểm trung tâm không thay đổi nhiều và từ đó hàm mất mát cũng sẽ ít bị ảnh hưởng. Vậy nên chúng ta có thể sử dụng 1 trong 3 cách trên để xác định điều kiện dừng của thuật toán

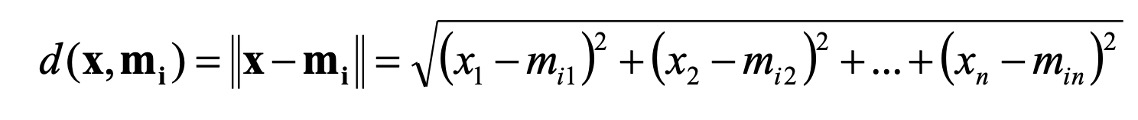
### Xác định điểm trung tâm của cluster

Để xác định điểm trung tâm của cluster ta sử dụng công thức như sau:  


Trong đó  là cluster thứ i, x là điểm trung tâm của cluster  tương ứng.

### Phép đo khoảng cách

Trong K-means để đánh giá mức độ giống nhau hay khoảng cách giữa 2 điểm dữ liệu ta có thể sử dụng các phép đo khoảng cách khác nhau. Ngoài khoảng cách Euclidean, tuỳ thuộc vào từng bài toán có thể sử dụng phương pháp đo khác (cosine, manhattan…)



Mọi phương pháp tính khoảng cách giữa 2 vector đều có thể được sử dụng. Mỗi cách tính khoảng cách thể hiện cách nhìn nhận về dữ liệu

Tuỳ vào từng đặc điểm của mỗi bài toán khác nhau, sẽ sử dụng những phương pháp đo khác nhau để phụ hợp với mỗi bài toán

## Một số yếu tố ảnh hưởng đến thuật toán Kmeans

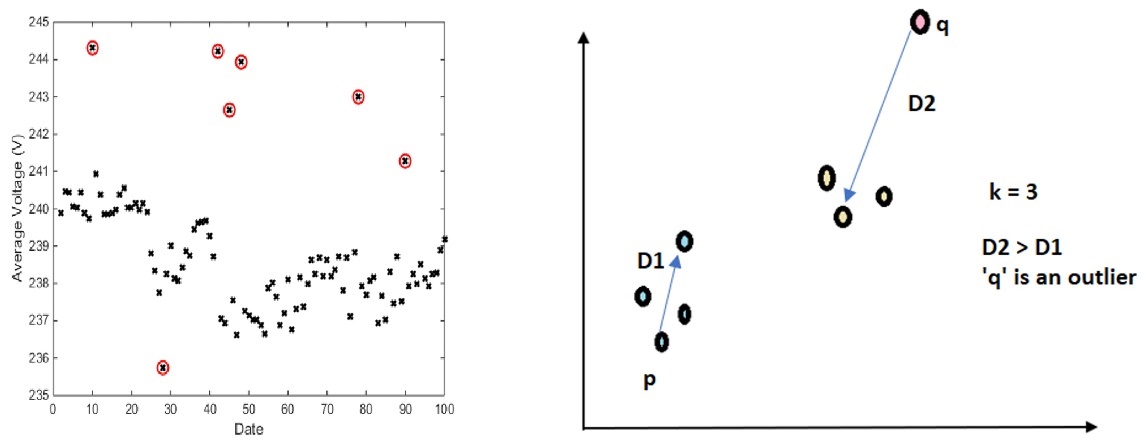
### Ảnh hưởng của outlier

Outliers là một hoặc nhiều cá thể khác hẳn đối với các thành viên còn lại của nhóm. Sự khác biệt này có thể dựa trên nhiều tiêu chí khác nhau như giá trị hay thuộc tính. Outlier có thể như là nhiễu trong các cảm biến hay lỗi trong quá trình nhập liệu của người dùng ảnh hưởng đến chất lượng của dữ liệu.

Xem xét ảnh hường:

K-means nhạy cảm với các điểm outlier, ví dụ: Các điểm dữ liệu outlier ảnh hưởng lớn đến kết quả của việc phân cụm:

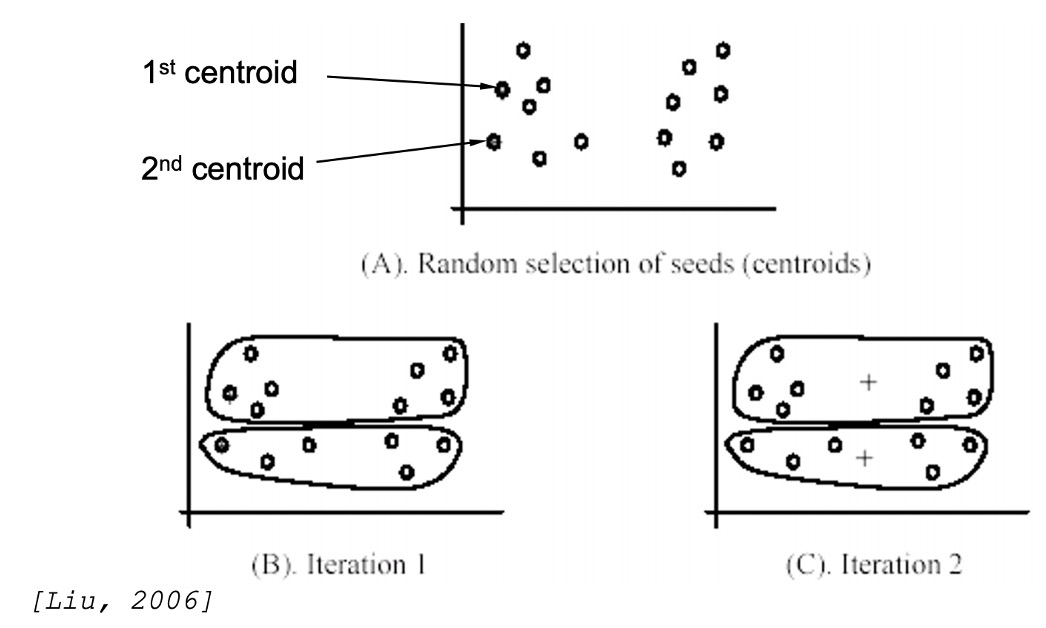
* Các điểm dữ liệu outlier có khoảng cách đến các điểm dữ liệu chuẩn rất lớn.
* Phân bố của các điểm outliner rất khác so với các điểm dữ liệu chuẩn
* Nhiễu hoặc lỗi của dữ liệu được thể hiện trong các điểm outlier



Khắc phục outlier:

* **Outlier removal**: Có thể loại bỏ các điểm dữ liệu xa đáng kể so với điểm trung tâm (centroid) của các cluster so với các điểm dữ liệu khác. Việc loại bỏ có thể được thực hiện trước hoặc trong khi phân cụm.
* **Random sampling**: Thay vì phân cụm toàn bộ tập dữ liệu, chúng ta sẽ lấy ngẫu nhiên tập con S từ tập dữ liệu huấn luyện. S được sử dụng để phân cụm, tập S lúc này sẽ có ít các điểm outlier hơn tập dữ liệu gốc. Sau khi phân cụm xong, tập dữ liệu còn lại sẽ được gán vào các cụm đã học được

### Ảnh hưởng của việc khởi tạo trung tâm

Chất lượng của K-means phụ thuộc vào việc khởi tạo các điểm centroid  


**Giải pháp 1**: Lặp lại nhiều lần thuật toán K-means:

* Mỗi lần chạy lại thuật toán K-means sẽ khởi tạo các điểm centroid khác nhau
* Sau quá trình học, tiến hành gộp các kết quả từ các lần chạy thành kết quả cuối cùng

**Giải pháp 2**: Thuật toán K-means++ : Để tìm ra cụm tốt nhất, chúng ta có thể lần lượt khởi tạo các điểm trung tâm

## Mô hình song song hóa

A diagram of a model

Description automatically generated

## Xây dựng mô hình Kmeans với song song hóa giải thuật

### Xây dựng mô hình Kmeans

**Hàm Euclidean** dùng để tính khoảng cách giữa hai điểm trong không gian nhiều chiều. Sử dụng thư viện NumPy để thực hiện các phép toán mảng nhanh chóng.

**Hàm assign\_to\_cluster** nhận một điểm dữ liệu point và danh sách các trung tâm của các cụm đã được broadcast trong Spark (sử dụng broadcast\_centers).

Đối với mỗi trung tâm của mỗi cụm, tính khoảng cách Euclidean giữa điểm dữ liệu và trung tâm đó bằng cách gọi hàm euclidean\_distance.

Sau đó, nếu khoảng cách nhỏ hơn khoảng cách nhỏ nhất hiện tại (min\_distance), tiến hành cập nhật min\_distance và chỉ số của cụm được gán (assigned\_cluster). Cuối cùng sẽ trả về chỉ số cụm được gán cho điểm dữ liệu đó.

Đoạn code đang sử dụng broadcast để chia sẻ giá trị của các trung tâm cụm giữa các máy trong môi trường Apache Spark, giúp tăng hiệu suất khi thực hiện các phép toán trên các RDD (Resilient Distributed Datasets).

A screen shot of a computer code

Description automatically generated

**Hàm kmeans\_clustering**

* + Đầu tiên tiến hành trích xuất dữ liệu từ cột features\_col của DataFrame và chuyển đổi nó thành một RDD.
  + Dùng hàm takeSample để chọn ngẫu nhiên k trung tâm từ dữ liệu đặc trưng và sau đó sử dụng sc.broadcast để broadcast danh sách trung tâm cụm đó.
  + Tiếp theo thực hiện vòng lặp để gán mỗi điểm dữ liệu vào cụm gần nhất và tính toán lại trung tâm mới cho cụm dữ liệu. Cuối cùng tiến hành cập nhật broadcast với các trung tâm mới và gán các trung tâm cuối cùng cho mỗi điểm dữ liệu.
  + Cuối cùng kmeans\_clustering sẽ trả về RDD

A screenshot of a computer code

Description automatically generated

### Đánh giá mô hình

Sử dụng độ đo Silhouette.

Giả sử mạng lưới được chia thành k cụm. Với mỗi node i, đặt:

a(i) là khoảng cách trung bình từ i tới tất cả các node trong cùng cụm với i.

b(i) là khoảng cách trung bình ngắn nhất từ i tới bất kỳ cụm nào không chứa i. Cụm tương ứng với b(i) này được gọi là cụm hàng xóm của i.

Khi đó:

*s*(i) nằm trong đoạn [-1,1].

*s*(i) càng gần 1 thì node i càng phù hợp với cụm mà nó được phân vào.

*s*(i) = 0 thì không thể xác định được i nên thuộc về cụm nào giữa cụm hiện tại và cụm hàng xóm của nó.

*s*(i) càng gần -1 thì chứng tỏ i bị phân sai cụm, nó nên thuộc về cụm hàng xóm chứ không phải cụm hiện tại.

Đối với mỗi k (số cụm) từ 2 đến 10 tiến hành xuất ra độ đo silhouette để so sánh. Nhận thấy rằng với k = 2 thì silhouette cao nhất (0.922). Đây là mô hình cho ra kết quả tốt nhất.

A screenshot of a computer program

Description automatically generated

# CHƯƠNG 5: KẾT QUẢ VÀ HƯỚNG PHÁT TRIỂN



## KẾT QUẢ VÀ KẾT LUẬN

### RFM Model (Recency - Frequency - Monetary Model)

RFM là một phương pháp được sử dụng để phân tích giá trị khách hàng. Nó thường được sử dụng trong marketing cơ sở dữ liệu (kiểu như dựa vào dữ liệu về khách hàng để tiếp thị sản phẩm) và marketing trực tiếp và đã nhận được sự chú ý đặc biệt trong ngành bán lẻ và dịch vụ.”

RFM định lượng giá trị của một khách hàng dựa trên 3 thông tin chính:

* Recency: Khoảng thời gian mua hàng gần đây nhất là bao lâu. Cho biết khách hàng có đang thực sự hoạt động gần thời điểm đánh giá. Chỉ số này càng lớn càng cho thấy xu hướng rời bỏ của khách hàng càng cao. Đó là một cảnh báo cho doanh nghiệp nên thay đổi sản phẩm để đáp ứng thị hiếu khách hàng hoặc thay đổi chính sách để nâng cao chất lượng phục vụ.
* Frequency: Tần suất mua hàng của khách hàng. Nếu khách hàng mua càng nhiều đơn thì giá trị về doanh số mang lại cho công ty càng cao và tất nhiên giá trị của họ càng lớn. Tuy nhiên nếu chỉ xét dựa trên tần suất mua hàng thì cũng chưa đánh giá được đầy đủ mức độ tác động lên doanh thu bởi bên cạnh đó, giá trị đơn hàng cũng là yếu tố trực tiếp cho thấy khách hàng tiềm năng như thế nào.
* Monetary: Là số tiền chi tiêu của khách hàng. Đây là yếu tố trực quan nhất ảnh hưởng tới doanh số. Hay nói cách khác, doanh nghiệp quan tâm nhất là khách hàng đã dành bao nhiêu tiền để mua sắm sản phẩm của công ty? Monetary sẽ tác động trực tiếp tới doanh thu và bị tác động gián tiếp thông qua 2 yếu tố còn lại là Recency và Frequency.

Sau khi mapping giá trị các trường Recency, Frequency, Monetary với rank tương ứng trong ngưỡng ranks là 10. Ta có phân phối điểm rank của khách hàng như thế nào thông qua biểu đồ histogram với số lượng bins = 10.

A screenshot of a computer

Description automatically generated

A graph of a number of blue squares

Description automatically generated

* Phân phối của 'Final Rank' cho thấy đa số khách hàng của công ty có điểm rank từ thấp đến trung bình (khoảng từ 2 đến 6), với số lượng lớn nhất tập trung ở rank 4 và 5. Điều này cho thấy phần lớn khách hàng có mức độ tương tác và giá trị giao dịch ở mức trung bình.
* Số lượng khách hàng giảm dần ở các rank cao (từ 7 trở lên), chỉ ra rằng có ít khách hàng đạt được điểm cao trong cả ba chỉ số RFM. Điều này có thể phản ánh một cơ hội để tăng cường mức độ tương tác và giá trị đối với những khách hàng này.
* Tại các rank thấp (1 và 2), có một số lượng đáng kể khách hàng, có thể là những người mới hoặc những khách hàng không thường xuyên mua sắm hoặc có giá trị mua sắm thấp.

Điều này cho thấy tập khách hàng của công ty đa phần sẽ nằm ở những điểm rank nằm trong khoảng trung bình, chẳng hạn từ 4-6. Với các điểm rank quá cao hoặc quá thấp thì số lượng khách hàng tập trung càng thấp.

Dựa vào biểu đồ, ta cũng có thể phân tập khách hàng thành 3 nhóm:

* Khách hàng ít tiềm năng - nhãn Low: Rank từ [0, 4)
* Khách hàng thông thường - Normal: Rank từ [4, 7)
* Khách hàng VIP - nhãn VIP: Rank [7, 9]

**Thống kê số lượng khách hàng theo mỗi Segment**

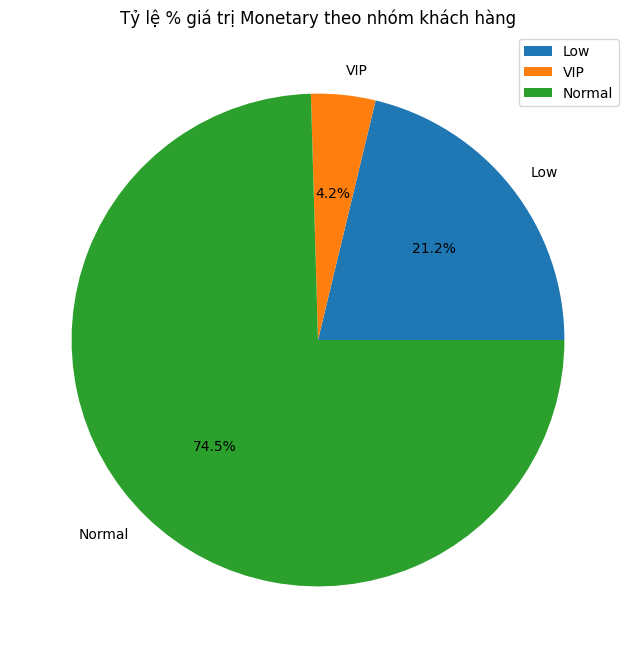
A pie chart with text on it

Description automatically generated

* Khách hàng Normal chiếm tỷ lệ lớn nhất trong cơ sở dữ liệu khách hàng với 64.1%, điều này cho thấy đa số khách hàng của công ty đang hoạt động ở mức độ tương tác trung bình.
* Khách hàng VIP chiếm 26.4% tổng số khách hàng, phản ánh một phần quan trọng của cơ sở khách hàng có giá trị cao đối với công ty.
* Khách hàng Low chỉ chiếm 9.5%, điều này cho thấy rằng nhóm khách hàng ít tiềm năng hoặc mới là phần nhỏ nhất của cơ sở khách hàng.

Những thông tin này hữu ích cho việc định hình chiến lược marketing và quản lý mối quan hệ khách hàng:

* Có thể tập trung vào việc nâng cấp khách hàng từ nhóm Normal lên VIP thông qua các chương trình khuyến mãi hoặc ưu đãi đặc biệt.
* Cải thiện sự hài lòng và mức độ gắn bó của khách hàng Low, có thể thông qua việc cung cấp trải nghiệm khách hàng cá nhân hóa hoặc chăm sóc khách hàng tốt hơn.
* Duy trì mối quan hệ với khách hàng VIP và tìm cách gia tăng giá trị qua mỗi giao dịch của họ, đồng thời khám phá các cơ hội để tăng số lượng khách hàng trong nhóm này.



* Khách hàng Normal: Chiếm phần lớn giá trị Monetary với 74.5%, điều này cho thấy rằng dù có số lượng lớn, khách hàng trong phân đoạn này có thể không chi tiêu nhiều như khách hàng VIP, nhưng tổng giá trị góp vào doanh số lại rất đáng kể.
* Khách hàng VIP: Tuy chỉ chiếm 21.2% của tổng giá trị Monetary, nhưng tỷ lệ này vẫn đáng chú ý, phản ánh rằng mặc dù số lượng khách hàng trong phân đoạn này có thể ít hơn, nhưng mỗi người trong họ có giá trị giao dịch cao, đóng góp lớn vào doanh thu.
* Khách hàng Low: Chỉ chiếm một tỷ lệ nhỏ là 4.2%, phản ánh rằng mặc dù họ mua sắm thường xuyên (như được thấy trong biểu đồ về tần suất mua hàng), nhưng giá trị giao dịch của họ thấp.

Nhận xét này cung cấp thông tin quan trọng cho việc lập kế hoạch chiến lược. Mặc dù khách hàng Normal có thể không chi tiêu nhiều như VIP trong mỗi giao dịch, tổng giá trị giao dịch của họ lại rất cao do số lượng lớn. Khách hàng VIP, với mức đóng góp cao cho doanh thu dù chiếm tỷ lệ thấp trong cơ sở khách hàng, cần được chăm sóc đặc biệt để duy trì và phát triển. Còn khách hàng Low có thể được xem là cơ hội tiềm năng để nâng cao giá trị giao dịch thông qua các chương trình khuyến mãi hoặc cải thiện trải nghiệm khách hàng.

A graph with blue rectangles

Description automatically generated

* **Khách hàng Low**: Mặc dù được gọi là phân đoạn "Low", nhóm này lại có doanh số trung bình cao nhất, với giá trị lên đến hơn 4500. Điều này có thể chỉ ra rằng những khách hàng này mua sắm ít lần nhưng mỗi lần mua lại có giá trị lớn, hoặc có thể là những khách hàng lớn nhưng không thường xuyên mua hàng.
* **Khách hàng VIP**: Phân đoạn này có doanh số trung bình thấp nhất, chỉ với khoảng 327. Điều này có thể phản ánh rằng dù những khách hàng này được xếp vào phân đoạn VIP dựa trên tần suất mua hàng và gần đây nhất họ đã mua hàng, nhưng từng đơn hàng của họ lại có giá trị thấp.
* **Khách hàng Normal**: Nhóm này có doanh số trung bình ở mức vừa phải, khoảng 2385, điều này cho thấy đây là nhóm khách hàng mua sắm đều đặn với giá trị mua hàng trung bình.

A graph with blue rectangles

Description automatically generated

* Khách hàng Low: Có tần suất mua hàng trung bình cao nhất. Điều này có thể cho thấy rằng, dù họ có thứ hạng thấp trong các chỉ số khác, họ vẫn mua sắm thường xuyên hơn so với hai nhóm còn lại.
* Khách hàng VIP: Tần suất mua hàng trung bình của họ thấp hơn so với phân đoạn Low, điều này có thể gây ngạc nhiên vì khách hàng VIP thường được kỳ vọng là những người mua sắm thường xuyên hơn. Điều này có thể chỉ ra rằng tiêu chí phân loại VIP cần được xem xét lại, hoặc VIP có thể được đánh giá cao vì các yếu tố khác như giá trị mua hàng lớn hơn dù họ mua ít lần hơn.
* Khách hàng Normal: Có tần suất mua hàng trung bình thấp hơn so với phân đoạn Low nhưng cao hơn phân đoạn VIP. Điều này có thể phản ánh hành vi mua sắm đều đặn nhưng không thường xuyên như khách hàng Low.

**Kết luận:**

* Chúng ta cần thay đổi chính sách về sản phẩm để chúng trở nên đa dạng và hấp dẫn khách hàng VIP hơn.
* Gia tăng số lượng khách hàng VIP để đảm bảo mục tiêu 20% khách hàng là VIP mang lại 80% doanh số.
* Đưa ra các chiến lược chăm sóc và giữ chân khách hàng hiệu quả.

### Kmeans Clustering

Chạy thuật toán kmeans với k (số cụm) là 2

A computer code with text

Description automatically generated with medium confidence

Ta được kết quả phân cụm và tiến hành join với dataframe ban đầu.

A screenshot of a computer code

Description automatically generated

Cuối cùng ta có được kết quả phân cụm như sau

A screenshot of a computer

Description automatically generated

Kết quả các cụm theo giá trị RFM Model

A bar graph with a bar and a cluster

Description automatically generated

A bar chart with a bar graph

Description automatically generated with medium confidence

A bar graph with orange and blue bars

Description automatically generated

**Các biểu đồ cho thấy mỗi cụm có các đặc điểm khác nhau về các biến RFM:**

* **Recency (Số ngày mua hàng gần nhất)**: Cụm 0 có giá trị Recency cao, điều này cho thấy khách hàng trong cụm này đã lâu không mua hàng. Đây có thể là những khách hàng "ngủ đông" hoặc không còn hoạt động. Họ cần được chú ý đặc biệt thông qua các chiến dịch tái kích hoạt hoặc giữ chân.
* **Frequency (Tần suất mua hàng)**: Cụm 1 lại cho thấy tần suất mua hàng cao, biểu thị một nhóm khách hàng trung thành, mua hàng thường xuyên. Đây là những khách hàng cốt lõi, có khả năng tạo ra đa số doanh thu qua các giao dịch lặp đi lặp lại.
* **Monetary (Giá trị tiền tệ)**: Cụm 1 cũng có giá trị Monetary cao, có nghĩa là nhóm này chi tiêu nhiều hơn. Đây có thể là những khách hàng chiến lược, đóng góp một phần lớn vào doanh số tổng thể.

**Nhận xét:**

* Cụm 0 có thể gồm những khách hàng đã từng mua sắm nhưng không gần đây, có thể do họ không hài lòng với sản phẩm/dịch vụ, hoặc họ có thể chỉ là những khách hàng không thường xuyên. Họ cần được thúc đẩy để tăng tần suất mua hàng.
* Cụm 1 xuất hiện như là nhóm khách hàng "vàng" - những người không chỉ mua sắm thường xuyên mà còn chi tiêu nhiều tiền. Những khách hàng này cần được nuôi dưỡng và duy trì mối quan hệ, cũng như cần được cung cấp các ưu đãi và trải nghiệm cá nhân hóa để tăng cường lòng trung thành và tối đa hóa giá trị của họ đối với công ty.

**Cách tiếp cận tương lai:**

1. **Tối ưu hóa phân khúc:** Sử dụng phân tích sâu hơn hoặc thuật toán phức tạp hơn để hiểu rõ hơn về hành vi và đặc điểm của từng cụm, từ đó tạo ra các chiến lược marketing và dịch vụ cá nhân hóa.
2. **Tương tác dựa trên dữ liệu:** Phát triển các chiến dịch tiếp thị dựa trên dữ liệu phân khúc để tăng cường tương tác và giữ chân khách hàng.
3. **Dự đoán giá trị khách hàng:** Áp dụng các mô hình dự đoán để ước lượng giá trị trọn đời của khách hàng (CLV) trong từng phân đoạn, hỗ trợ việc ra quyết định chiến lược.
4. Tự động hóa và tinh chỉnh: Sử dụng học máy và AI để tự động hóa quá trình phân khúc và tinh chỉnh các phân đoạn theo thời gian dựa trên hành vi khách hàng đang thay đổi.
5. Tích hợp với các nền tảng khác: Kết hợp dữ liệu RFM với dữ liệu từ các nền tảng khác như mạng xã hội, web analytics để có cái nhìn toàn diện hơn về khách **hàng.**

Kết luận so sánh hai phương pháp phân khúc khách hàng - sử dụng trung bình rank (Average Rank Method) và K-means Clustering - chúng ta có thể xem xét một số khía cạnh khác nhau:

1. **Độ Chính Xác của Phân Cụm:**
   * Average Rank Method: Phương pháp này dễ dàng để thực hiện và giải thích. Nó cung cấp một cái nhìn trực quan về việc phân bố của khách hàng dựa trên rank từ các chỉ số RFM. Tuy nhiên, phương pháp này chưa phản ánh chính xác mối quan hệ phức tạp giữa các biến hoặc nắm bắt được các hành vi khách hàng ẩn.
   * K-means Clustering: Phân cụm K-means có thể phát hiện các mô hình phức tạp hơn và tạo ra các nhóm khách hàng dựa trên sự tương đồng thực tế trong dữ liệu. Phương pháp này cần được lập số lượng cụm một cách cẩn thận và sử dụng kỹ thuật giảm chiều dữ liệu như PCA để hiệu quả hơn. Đã tạo ra các cụm dựa trên dữ liệu RFM và thêm vào đó là dữ liệu one-hot encoded của 'Country'. Các biểu đồ cho thấy rõ ràng sự khác biệt giữa các cụm về Recency, Frequency, và Monetary.
   * Ta thấy RFM đưa ra kết quả phân nhóm trên rank khó giải thích hơn, nhiều bất đồng hơn, không có sự khác biệt giữa các cụm về Recency, Frequency, và Monetary trong khi Kmeans đã làm được điều đó với chỉ số Silhouette score lên tới 0.9.

Quyết định sử dụng phương pháp nào cần dựa trên mục tiêu cụ thể của doanh nghiệp, nguồn lực có sẵn, và mức độ chi tiết thông tin mà bạn muốn thu được từ quá trình phân khúc.

## HƯỚNG PHÁT TRIỂN

Phân khúc khách hàng là một trong những nhiệm vụ quan trọng trong quản lý mối quan hệ với khách hàng (CRM) và marketing. Việc phân khúc có thể giúp công ty hiểu rõ hơn về cơ sở khách hàng của mình, từ đó phát triển các chiến lược tiếp cận và tối ưu hóa dịch vụ.

Thông qua phân khúc khách hàng, Chúng ta có thể phát triển các chiến lượt kinh doanh và bộ máy tổ chức doanh nghiệp phù hợp với mục tiêu phục vụ trên từng phân khúc. Đồng thời theo dõi chuyển dịch cơ cấu khách hàng trên từng phân khúc qua thời gian cũng giúp đánh giá được mức độ phát triển tập khách hàng của công ty ra sao? Công ty cần tung ra những điều chỉnh, chiến lược như thế nào để phát triển theo hướng gia tăng tỷ trọng khách hàng VIP, giữ chân khách hàng quay lại mua sắm thường xuyên hơn và nâng cao giá trị đơn hàng trên mỗi lượt mua sắm.

Dưới đây là một số hướng phát triển tương lai cho bài toán phân khúc khách hàng mà nhóm chúng em đã liệt kê ra:

**Phân tích RFM sâu hơn**: Tiếp tục phân tích RFM bằng cách tinh chỉnh các ngưỡng cho Recency, Frequency và Monetary để phân định rõ ràng hơn giữa các nhóm khách hàng, có thể bao gồm việc tạo ra các phân đoạn phụ để nhận diện những nhóm khách hàng đặc biệt như "khách hàng mới nổi" hoặc "khách hàng có nguy cơ rời bỏ".

**Các biển thế khác của RFM:**

* RFD - Recency, Frequency, Duration (thời gian) là phiên bản đã được modified của phân tích RFM. Nhưng thay vì phân tích giá trị khách hàng, RFD được sử dụng để phân tích hành vi khách hàng theo các nhóm người xem/người đọc/người lướt web.
* RFE - Recency, Frequency, Engagement(mức độ cam kết) là phiên bản mở rộng của phân tích RFD nhằm xác định mức độ gắn bó của khách hàng đối với một nền tảng web, app. Trường Engagement (mức độ cam kết) được xác định thông qua thời lượng truy cập, số trang trên mỗi lượt truy cập và các chiều dữ liệu khác tương tự. Mô hình RFE có thể được sử dụng để phân tích hành vi khách hàng theo các nhóm người xem/người đọc/người lướt web.
* RFM-I - Recency, Frequency, Monetary Value – Interactions (Giá trị tiền tệ - Tương tác) là một phiên bản khác của RFM để đánh giá chi phí tương tác marketing trong tiếp cận khách hàng.
* RFMTC - Recency, Frequency, Monetary Value, Time (Thời gian), Churn rate (Tỷ lệ rời bỏ) là một mô hình RFM mở rộng được đề xuất bởi I-Cheng và cộng sự(2009). Mô hình sử dụng chuỗi Bernoulli trong lý thuyết xác suất nhằm dự báo xác suất mua hàng tại những chiến dịch marketing tiếp theo.

**Áp dụng Machine Learning**: Sử dụng các thuật toán Machine Learning nâng cao như phân cụm cơ sở (base segmentation) dựa trên các đặc trưng khác nhau của khách hàng, không chỉ dừng lại ở RFM. Các thuật toán như hierarchical clustering, DBSCAN, hoặc các mô hình phân loại có giám sát như SVM, random forest có thể được áp dụng để phân khúc khách hàng một cách tự động và phức tạp hơn.

**Tích hợp Big Data:** Khai thác Big Data và các nguồn dữ liệu bổ sung như dữ liệu từ truyền thông xã hội, hành vi trực tuyến, lịch sử mua hàng để có cái nhìn đa chiều và toàn diện hơn về khách hàng.

**Personalization và Micro-segmentation:** Chuyển từ phân khúc rộng sang micro-segmentation, nơi mỗi khách hàng hoặc nhóm nhỏ khách hàng có các chiến lược tiếp cận riêng biệt dựa trên hành vi và sở thích cá nhân của họ.

**Tích hợp AI để dự đoán và tự động hóa:** Sử dụng AI để dự đoán xu hướng mua hàng trong tương lai và tự động hóa việc gửi thông điệp marketing, ưu đãi cá nhân hóa dựa trên phân khúc khách hàng.

**Dynamic Segmentation:** Phát triển các mô hình có khả năng điều chỉnh phân khúc một cách động dựa trên hành vi gần đây của khách hàng, cho phép phản ứng nhanh chóng với sự thay đổi trong nhu cầu và hành vi của họ.

**Tích hợp với Omnichannel Marketing**: Đảm bảo rằng phân khúc khách hàng được tích hợp mượt mà với các chiến lược marketing đa kênh, từ offline đến online, để tạo ra trải nghiệm khách hàng nhất quán và cá nhân hóa.

**Phân tích tâm lý khách hàng**: Sử dụng tâm lý học tiêu dùng và dữ liệu phân tích hành vi để hiểu sâu hơn về động cơ và nguyện vọng của khách hàng, từ đó phát triển sản phẩm và dịch vụ phù hợp.

Mỗi hướng phát triển đều cần được xây dựng trên nền tảng dữ liệu chất lượng cao và hiểu biết sâu sắc về khách hàng. Đồng thời, việc liên tục đánh giá và điều chỉnh chiến lược dựa trên phản hồi và kết quả thu được là rất quan trọng để đảm bảo hiệu quả của việc phân khúc.

# CHƯƠNG 6: THAM KHẢO

1. [Áp dụng model Machine learning vào bài toán phân khúc khách hàng](https://viblo.asia/p/ap-dung-model-machine-learning-vao-bai-toan-phan-khuc-khach-hang-Qbq5QjzEKD8)
2. [RMF model wikipedia](https://en.wikipedia.org/wiki/RFM_(customer_value))
3. [RMF model with σ-scaled hadron masses and couplings for description of heavy-ion collisions below 2A GeV](https://arxiv.org/abs/1902.09016)
4. [Using RFM to Identify Your Best Customers](https://www.eightleaves.com/2011/01/using-rfm-to-identify-your-best-customers/)
5. [Making Your Database Pay Off Using Recency Frequency and Monetary Analysis](http://www.dbmarketing.com/2010/03/making-your-database-pay-off-using-recency-frequency-and-monetary-analysis/)