**Media Campaign Cost**

**Abstract:**

Dự án này nhằm mục đích dự đoán chi phí của các chiến dịch truyền thông tại các cửa hàng thực phẩm sử dụng các mô hình học máy. Chúng tôi đã sử dụng các phương pháp chính là GradientBoostingRegressor, RandomForestRegressor, và XGBRegressor để phát triển các mô hình dự đoán dựa trên các đặc trưng được cung cấp trong bộ dữ liệu. Random Forest Regressor tổng hợp sức mạnh của nhiều cây quyết định để tạo ra một mô hình mạnh mẽ và linh hoạt, trong khi Gradient Boosting Regressor xây dựng các mô hình tuần tự tập trung vào việc sửa lỗi từ mô hình trước, và XGBoost cung cấp sự linh hoạt và hiệu suất tốt trên các tập dữ liệu lớn. Kết quả của chúng tôi cho thấy cả ba mô hình đều cung cấp các dự đoán chi phí chính xác cho các chiến dịch truyền thông. Dự án này nhấn mạnh tầm quan trọng của học máy trong việc tối ưu hóa các chiến lược tiếp thị và phân bổ ngân sách cho các chiến dịch truyền thông trong môi trường bán lẻ.

**Keywords:**

DANH MỤC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| **Từ viết tắt** | **Từ nguyên gốc** |
| GBRegressor | GradientBoostingRegressor |
| RFRegressor | RandomForestRegressor |
| XGBRegressor | eXtreme Gradient Boosting |

1. **Introduction**

## Đông lực nghiên cứu

Nghiên cứu về chi phí cho chiến dịch truyền thông không chỉ là việc đánh giá số liệu, mà còn là việc hiểu rõ những yếu tố chi phí ẩn sau sự hiện diện độc đáo của mỗi chiến dịch truyền thông. Dưới đây là những động lực chính đằng sau quá trình tìm hiểu về chi phí của chiến dịch truyền thông:

* Tối Ưu Hóa Ngân Sách:

Nắm vững chi phí giúp tối ưu hóa ngân sách, một yếu tố quyết định sự thành công của chiến dịch. Nghiên cứu này sẽ giúp định hình chiến lược ngân sách phù hợp với mục tiêu và phạm vi mong muốn.

* Hiểu Rõ Cơ Cấu Chi Phí:

Phân tích chi tiết cơ cấu chi phí bao gồm chi phí quảng cáo, sản xuất nội dung, và phân phối, giúp hiểu rõ hơn về nguồn gốc của mỗi khoản chi phí và cách chúng ảnh hưởng đến kết quả.

* Tối Ưu Hóa Hiệu Suất:

Nghiên cứu giúp xác định những chiến lược chi phí mang lại hiệu suất tốt nhất, từ việc chọn nền tảng đến định hình thông điệp quảng cáo.

* Ưu Tiên Chi Phí và Tác Động Xã Hội:

Đánh giá tác động của chi phí đối với mục tiêu quảng cáo và xã hội. Có thể xem xét chiến lược giảm giá, ưu đãi cho các tổ chức phi lợi nhuận, hoặc chiến dịch có tác động tích cực đối với cộng đồng.

* Tìm Kiếm Cơ Hội Tăng Cường Giá Trị:

Nghiên cứu giúp xác định cơ hội để tăng cường giá trị từ mỗi đồng chi phí, bằng cách sáng tạo nội dung sáng tạo và tối ưu hóa chiến lược quảng bá.

* Đối Mặt với Thách Thức Tăng Chi Phí:

Nghiên cứu sẽ làm rõ những thách thức và nguy cơ tăng chi phí và đề xuất chiến lược giải quyết để duy trì tính khả thi của chiến dịch.

Qua quá trình nghiên cứu này, chúng ta sẽ không chỉ hiểu rõ về chi phí mà còn đặt ra cơ hội để tối ưu hóa chiến dịch truyền thông, tạo ra những ảnh hưởng tích cực và bền vững trong thị trường đầy cạnh tranh ngày nay.

## **Mục tiêu và phạm vi nghiên cứu**

**1.2.1. Mục tiêu**

Ở nghiên cứu này, nhóm chúng tôi ưu tiên các mục tiêu sau:

* Thứ nhất, thành thạo ngôn ngữ lập trình Python và cách thức train một model học máy.
* Thứ hai, nghiên cứu và thử nghiệm các phương pháp trên bộ dữ liệu
* Thứ ba, đưa ra các nhận định và ý kiến về bộ dữ liệu và phương pháp.
* Thứ tư, nhận xét về các phương pháp đã train.

**1.2.2. Phạm Vi**

Nghiên cứu được triển khai dựa trên bộ dữ liệu sẵn có trên Kaggle.

## 1.3. Đóng góp của nghiên cứu

Nghiên cứu về chi phí cho chiến dịch truyền thông có thể mang lại nhiều đóng góp quan trọng cho ngành truyền thông, quảng cáo và kinh doanh nói chung. Dưới đây là một số đóng góp tiêu biểu mà nghiên cứu của bạn có thể mang lại:

1. Tối Ưu Hóa Ngân Sách:

Cung cấp các chiến lược tối ưu hóa ngân sách để doanh nghiệp và quảng cáo gia tăng hiệu quả chi phí trong chiến dịch truyền thông của họ.

1. Hiểu Rõ Hơn về Chi Phí Quảng Cáo:

Đưa ra cái nhìn chi tiết về cơ cấu chi phí trong quảng cáo, giúp doanh nghiệp hiểu rõ hơn về nơi mà chi phí được tiêu và tìm cách để tối ưu hóa chúng.

1. Tạo Ra Chiến Lược Chi Phí Bền Vững:

Phát triển các chiến lược chi phí có tính bền vững, giúp doanh nghiệp duy trì và cải thiện hiệu suất chiến dịch truyền thông theo thời gian.

1. Hỗ Trợ Quyết Định Chiến Lược:

Cung cấp thông tin chi tiết để hỗ trợ quyết định chiến lược về nền tảng quảng cáo, đối tượng mục tiêu, và thời điểm phát sóng, tối ưu hóa ảnh hưởng và chi phí.

1. Phản Hồi Đối với Xã Hội:

Đánh giá tác động xã hội của chiến dịch truyền thông thông qua việc đưa ra các đề xuất và chiến lược có thể hỗ trợ cộng đồng và góp phần vào mục tiêu xã hội.

1. Ứng Dụng Cho Các Lĩnh Vực Khác Nhau:

Nghiên cứu có thể cung cấp các hướng dẫn và chiến lược áp dụng cho nhiều lĩnh vực, từ doanh nghiệp nhỏ đến các tổ chức lớn và từ các chiến dịch quảng cáo thương mại đến các chiến dịch xã hội.

1. Giảm Rủi Ro và Thách Thức:

Đưa ra phân tích về rủi ro và thách thức có thể phát sinh trong việc quản lý chi phí chiến dịch truyền thông, giúp doanh nghiệp chuẩn bị và đối mặt với những thách thức này một cách hiệu quả.

1. Khuyến Khích Sự Sáng Tạo:

Khuyến khích sự sáng tạo trong quảng cáo và truyền thông thông qua việc đề xuất chiến lược chi phí linh hoạt và đáp ứng đa dạng nhu cầu của thị trường.

1. Xác Định Cơ Hội Tăng Cường Giá Trị:

Phân loại cơ hội để tăng cường giá trị từ chiến dịch truyền thông thông qua việc tối ưu hóa nội dung và phương tiện truyền thông.

Những đóng góp này có thể giúp cải thiện chiến lược quảng cáo và truyền thông, làm tăng cường hiệu quả và giảm chi phí, đồng thời đảm bảo tác động tích cực đối với cộng đồng và xã hội.

1. **Related Works**

# Machine learning in marketing: A literature review, conceptual framework, and research agenda

Trong những năm gần đây, học máy (ML) và trí tuệ nhân tạo (AI) đã thu hút sự chú ý đáng kể trong các lĩnh vực công nghiệp khác nhau, bao gồm cả lĩnh vực marketing. ML và AI hứa hẹn mang lại cơ hội lớn cho việc làm cho marketing thông minh và hiệu quả hơn. Trong nghiên cứu này, chúng tôi tiến hành một đánh giá về tài liệu của các nghiên cứu trong các tạp chí học thuật về ứng dụng ML trong marketing và đề xuất một khung khái niệm nhấn mạnh các công cụ và công nghệ ML chính là nền tảng của các ứng dụng ML trong marketing. Chúng tôi sử dụng 7Ps của marketing mix, bao gồm sản phẩm, giá cả, khuyến mãi, địa điểm, người, quy trình và bằng chứng vật lý, để phân tích các ứng dụng này từ 140 bài viết được lựa chọn. Các ứng dụng được hỗ trợ bởi các công cụ ML khác nhau (phân tích văn bản, giọng nói, hình ảnh và video) và các kỹ thuật như giám sát, không giám sát và thuật toán học bổng. Chúng tôi đề xuất một khung khái niệm hai lớp cho việc phát triển các ứng dụng ML trong marketing. Khung này có thể phục vụ cho nghiên cứu tương lai và cung cấp một minh họa về sự phát triển của các ứng dụng ML trong marketing.

* **Machine learning and artifcial intelligence use in marketing: a general taxonomy**

Sự xuất hiện của dữ liệu được tạo ra bởi người tiêu dùng và sự phổ biến ngày càng tăng của các kỹ thuật Học Máy (ML) đang làm thay đổi cách thức thực hiện marketing. Các nhà tiếp thị và nhà nghiên cứu vẫn chưa có hiểu biết sâu rộng về loạt các cơ hội mà các ứng dụng ML mang lại trong việc tạo và duy trì lợi thế cạnh tranh cho doanh nghiệp. Trong bài báo này, chúng tôi đề xuất một phân loại về các trường hợp sử dụng ML trong marketing dựa trên một đánh giá có hệ thống từ văn học học thuật và doanh nghiệp. Chúng tôi đã xác định 11 trường hợp sử dụng lặp lại, được tổ chức trong 4 gia đình đồng nhất tương ứng với các lĩnh vực cơ bản của ML trong marketing, bao gồm: các nguyên tắc cơ bản của người mua, trải nghiệm tiêu thụ, ra quyết định và tác động tài chính. Chúng tôi thảo luận về các mẫu lặp lại được xác định trong phân loại và cung cấp một khung khái niệm cho việc diễn giải và mở rộng nó, nhấn mạnh các tác động thực tiễn đối với các nhà tiếp thị và nhà nghiên cứu.

# Machine Learning in Marketing: A Systematic Literature and Text Mining Research

Thảo luận về trí tuệ nhân tạo (AI) và học máy (ML) đã ngày càng trở nên quan trọng trong lý thuyết và thực hành marketing. Do đó, tình trạng hiện tại của nghiên cứu về marketing liên quan đến AI và ML đang phong phú với các ý tưởng. Thật không may, luồng nghiên cứu này có cấu trúc yếu và thiếu một đánh giá có hệ thống về văn bản hiện có trong lĩnh vực nghiên cứu. Để điền vào khoảng trống này, bài báo này nhằm mục đích trả lời hai câu hỏi nghiên cứu sau: (1) Cơ thể kiến thức được cấu trúc như thế nào đối với việc ứng dụng ML trong nghiên cứu marketing? và (2) Các khái niệm ML và các chủ đề marketing điển hình nào được xem xét trong nghiên cứu hiện tại? Để trả lời những câu hỏi này, chúng tôi tiến hành một đánh giá có hệ thống về văn bản hiện có bằng cách áp dụng ML (khai thác văn bản) trên khoảng 900 bài báo. Kết quả đã phát hiện ra các thông tin quan trọng về cấu trúc và hướng của nghiên cứu hiện tại về ML trong lĩnh vực marketing, điều này có thể góp phần vào các dự án nghiên cứu tương lai và hướng dẫn các nhà tiếp thị trong lĩnh vực nghiên cứu quan trọng này.

# Machine learning and AI in marketing – Connecting computing power to human insights

Trong thế giới kinh doanh, các đại lý trí tuệ nhân tạo (AI) được điều khiển bởi các thuật toán học máy đang nhanh chóng biến đổi, thu hút sự quan tâm tăng cao từ các nhà nghiên cứu. Trong bài báo này, chúng tôi xem xét và kêu gọi cho nghiên cứu marketing sử dụng các phương pháp học máy. Chúng tôi cung cấp một tổng quan về các nhiệm vụ và phương pháp học máy phổ biến, và so sánh chúng với các phương pháp thống kê và kinh tế mà các nhà nghiên cứu marketing truyền thống thường sử dụng. Chúng tôi lập luận rằng các phương pháp học máy có thể xử lý dữ liệu quy mô lớn và không cấu trúc, và có cấu trúc mô hình linh hoạt mang lại hiệu suất dự đoán mạnh mẽ. Tuy nhiên, các phương pháp như vậy có thể thiếu tính minh bạch và khả năng diễn giải mô hình. Chúng tôi thảo luận về các xu hướng và thực hành công nghiệp trí tuệ nhân tạo đáng chú ý, và xem xét về văn học marketing học vẫn còn non nớt sử dụng các phương pháp học máy. Quan trọng hơn, chúng tôi trình bày một khung khái niệm thống nhất và một chương trình nghiên cứu đa chiều. Từ năm khía cạnh chính của nghiên cứu thực nghiệm về marketing: phương pháp, dữ liệu, sử dụng, vấn đề và lý thuyết, chúng tôi đề xuất một số ưu tiên nghiên cứu, bao gồm mở rộng các phương pháp học máy và sử dụng chúng như các thành phần cốt lõi trong nghiên cứu marketing, sử dụng các phương pháp này để trích xuất thông tin từ dữ liệu quy mô lớn không cấu trúc, theo dõi và dữ liệu mạng, sử dụng chúng theo cách minh bạch cho phân tích mô tả, nguyên nhân và khuyến nghị, sử dụng chúng để vẽ ra hành trình mua hàng của khách hàng và phát triển khả năng hỗ trợ quyết định, và kết nối các phương pháp với những hiểu biết của con người và các lý thuyết marketing. Cơ hội nhiều cho các phương pháp học máy trong marketing, và chúng tôi hy vọng chương trình nghiên cứu đa chiều của chúng tôi sẽ truyền cảm hứng cho thêm nhiều công việc trong lĩnh vực hấp dẫn này.

**3. Data Preparation**

Tên bộ dữ liệu : Media Campaign Cost Dataset.

Food Mart (CFM) là một chuỗi cửa hàng tiện lợi ở Hoa Kỳ. Trụ sở chính của công ty tư nhân được đặt tại Mentor, Ohio và hiện có khoảng 325 cửa hàng được đặt tại Mỹ. Chợ Thực Phẩm Tiện Lợi hoạt động theo hệ thống nhượng quyền.

Food Mart là chuỗi cửa hàng tiện lợi lớn thứ ba toàn quốc tính đến năm 1988.

Sàn giao dịch NASDAQ đã loại bỏ Siêu thị Thực phẩm Tiện lợi cùng năm khi công ty không đáp ứng được yêu cầu báo cáo tài chính.

Carden & Cherry quảng cáo Chợ Thực phẩm Tiện lợi với nhân vật Ernest vào những năm 1980.

Nhiệm vụ của Chúng tôi là đưa ra Mô hình học máy giúp dự đoán chi phí của các chiến dịch truyền thông tại các siêu thị ẩm thực dựa trên các tính năng được cung cấp.

Mô tả tập dữ liệu:

store\_sale (tính bằng triệu): tổng tiền cửa hàng bán được

unit\_sale(tính bằng triệu): tổng tiền mặt hàng bán được

Total\_children: Tổng số trẻ

avg\_cars\_at home (xấp xỉ): trung bình xe ô tô tại nhà

Num\_children\_at\_home - num\_children\_at\_home: số trẻ ở nhà

Gross\_weight: trọng lượng mặt hàng

Recyclable\_package: có tái chế được không

Low\_fat : mặt hang có ít béo không

Units\_per\_case : số hộp trên kệ từng cửa hàng

Store\_sqft: diện tích cửa hàng

Coffee\_bar: số bar café có sẵn tại cửa hàng

Video\_store: Cửa hàng video/game có sẵn

Salad\_bar: Số quầy salad có sẵn tại cửa hàng

Prepare\_food: thực phẩm được chuẩn bị sẵn trong cửa hàng

Flowers store: kệ hoa có sẵn tại cửa hàng

Cost - CHI PHÍ THU NHẬP KHÁCH HÀNG tính bằng đô la

| **Dataset** | **Features** | **Source** |
| --- | --- | --- |
| train\_dataset.csv | store\_sales(in millions), unit\_sale, total\_children, num\_children\_at\_home, avg\_cars\_at home, gross\_weight, recyclable\_package, low\_fat, units\_per\_case, store\_sqft, coffee\_bar, video\_store, Salad\_bar, Prepare\_food, Flowers store, Cost | Kaggle.com |

Đây là dữ liệu đã được trực quan hóa bằng thư viện matplotlib dữ liệu gồm có 16 trường truong đó có 7 trường ở dạng Boolean (có giá trị trả về là false và true ứng với 0 và 1) là recyclable\_package, low\_fat, coffee\_bar, video\_store, salad\_bar, prepared food, florist.

Có thể thấy ở trường store\_sales số lượng cửa hàng phần lớn kinh doanh được 1-9 triệu đô. Số sản phẩm trong kệ cũng nằm trong khoảng 2-4 triệu sản phẩm và nhiều nhất là khoảng 3 triệu sản phẩm.

Số trẻ ở nhà phần lớn là 0 tức phần lớn là chưa lập gia đình hoặc là gia đình 1 vợ 1 chồng không có con nhỏ.

Số lương xe ở nhà phần lớn là 1- 3 chiếc. tỉ lệ giữa hàng tái chế, cửa hàng hoa, quầy thức ăn chế biến sẵn, quầy salad, quầy caffe và không có gần như là 1:1.

Trọng lượng của đồ vật được phân bố đều.

Chi phí mua sắm phải bỏ ra từ 50 đến 150 đô cho 1 lần mua sắm nhiều nhất là 50 và 75 đô còn lại phân bố không đồng đều.

Ảnh có chứa văn bản, biểu đồ, ảnh chụp màn hình, Sơ đồ

Mô tả được tạo tự động

**4. Methods**

## 4.1. Phương thức huấn luyện

Nhóm chúng em sẽ trình bày 3 phương pháp để huấn luyện mô hình là Gradient Boosting Regressor, Random Forest Regressor và XGBRegressor.

**Điểm giống:**

* Cả ba đều thuộc loại mô hình tập hợp, tức là chúng kết hợp dự đoán từ nhiều mô hình yếu để tạo ra một mô hình mạnh mẽ hơn.
* Cả ba mô hình đều dựa trên cây quyết định như là các mô hình cơ sở. Các cây quyết định được sử dụng để xây dựng mô hình con trong cả hai trường hợp.
* Cả ba đều có khả năng xử lý nhiễu tốt và ít yêu cầu chuẩn bị dữ liệu phức tạp.
* Cả ba đều có các siêu tham số có thể được tinh chỉnh để tối ưu hóa hiệu suất trên dữ liệu cụ thể.

**Điểm Khác:**

* Cơ Chế Tạo Ra Các Cây:

**Random Forest** tạo ra nhiều cây quyết định ngẫu nhiên độc lập và kết hợp kết quả, trong khi Gradient Boosting và XGBoost tạo ra các cây tuần tự, mỗi cây tập trung vào việc sửa lỗi của cây trước đó.

* Cách Tối Ưu Hóa Hàm Mất Mát:

**Gradient Boosting** và **XGBoost** sử dụng phương pháp gradient descent để tối ưu hóa hàm mất mát. Trong khi đó, Random Forest không thực hiện tối ưu hóa trực tiếp trên hàm mất mát.

* Quản Lý Trọng Số (Weights):

Trong Gradient Boosting, mỗi cây có một trọng số được điều chỉnh để cân nhắc đối với dự đoán cuối cùng. Trong Random Forest, mỗi cây có trọng số bằng nhau. XGBoost cũng có cơ chế tương tự như Gradient Boosting.

* Xử Lý Dữ Liệu Categorical:

XGBoost có tính năng tích hợp để xử lý dữ liệu hạng mục (categorical) mà không cần phải chuyển đổi chúng thành số như Random Forest. Gradient Boosting thông thường yêu cầu xử lý trước dữ liệu hạng mục.

Bộ dữ liệu Media Campaign Cost có cả nhiều đặc trưng và có các đặc trưng tương tác nên nhóm quyết định dùng các phương pháp trên để train model.

## 4.2. Quá trình huấn luyện Mô Hình

### **Tổng quan về các phương pháp**

Để train model này nhóm đã sử dụng các phương pháp nêu trên là Gradient Boosting Regressor, Random Forest Regressor và XGBRegressor.

#### Random Forest Regressor

Random Forest Regressor là một mô hình học máy thuộc loại Ensemble Learning, được xây dựng trên cơ sở của cây quyết định (Decision Trees). Được giới thiệu bởi Leo Breiman vào năm 2001, Random Forest Regressor tổng hợp sức mạnh của nhiều cây quyết định để tạo ra một mô hình mạnh mẽ, linh hoạt và chống overfitting.

Dưới đây là chi tiết về cách Random Forest Regressor hoạt động:

1. Cây Quyết Định (Decision Trees):

Random Forest bắt đầu với việc xây dựng nhiều cây quyết định độc lập. Mỗi cây quyết định được xây dựng dựa trên một phần ngẫu nhiên của dữ liệu (bootstrap sample) và một phần ngẫu nhiên của các đặc trưng (feature).

Mỗi cây quyết định được xây dựng theo cách chia nhánh dựa trên sự phân chia tốt nhất của dữ liệu dựa trên một đặc trưng cụ thể tại mỗi nút.

2. Bagging (Bootstrap Aggregating):

Random Forest sử dụng phương pháp bagging để tạo ra các mẫu bootstrap, tức là tạo ra nhiều tập con dữ liệu từ tập dữ liệu huấn luyện bằng cách chọn ngẫu nhiên và có thay thế.

Mỗi cây quyết định được huấn luyện trên một tập con dữ liệu khác nhau, giúp giảm sự phụ thuộc giữa các cây.

3. Random Feature Selection:

Tại mỗi nút trong cây quyết định, chỉ một số lượng nhỏ đặc trưng được chọn ngẫu nhiên để đưa ra quyết định.

Việc này giúp mỗi cây quyết định trở nên đa dạng và độc lập, ngăn chặn việc overfitting.

4. Voting/Averaging:

Khi có một dự đoán mới, Random Forest kết hợp dự đoán của tất cả các cây quyết định. Đối với bài toán regression, thông thường được sử dụng trung bình cộng của dự đoán của từng cây.

5. Hiệu Suất và Điều Chỉnh Tham Số:

Random Forest thường cho hiệu suất tốt trên nhiều loại dữ liệu mà không cần nhiều điều chỉnh tham số.

Một số tham số quan trọng bao gồm số lượng cây (n\_estimators), độ sâu tối đa của cây (max\_depth), và số lượng đặc trưng được chọn ở mỗi nút (max\_features).

6. Ưu điểm và Nhược điểm:

Ưu điểm: Robust trước overfitting, hiệu suất tốt trên dữ liệu nhiễu, có thể xử lý nhiều loại đặc trưng.

Nhược điểm: Không dễ giải thích như một cây quyết định đơn lẻ, tốn nhiều tài nguyên tính toán.

Random Forest Regressor là một lựa chọn mạnh mẽ cho nhiều bài toán regression, đặc biệt là khi có sự phức tạp và đa dạng trong dữ liệu.

#### Gradient Boosting Regressor

Gradient Boosting là một kỹ thuật Ensemble Learning cũng giống như Random Forest, nhưng thay vì xây dựng nhiều mô hình độc lập, Gradient Boosting xây dựng các mô hình theo cách tuần tự và tập trung vào việc sửa lỗi của mô hình trước đó. Một trong những biến thể phổ biến nhất của Gradient Boosting là Gradient Boosting Machines (GBM). Dưới đây là chi tiết về cách Gradient Boosting hoạt động:

1. Cây Quyết Định Nhỏ (Decision Stumps):

Mô hình Gradient Boosting thường sử dụng các cây quyết định nhỏ, thường chỉ có một nút quyết định và hai lá.

Cây quyết định nhỏ được gọi là "decision stump" và được sử dụng để tạo ra dự đoán sơ bộ.

2. Tính Hiệu Quả (Residuals):

Mô hình bắt đầu với một dự đoán sơ bộ, thường là trung bình của các giá trị đầu ra thực tế.

Tính hiệu quả (residuals) là sự chênh lệch giữa dự đoán sơ bộ và giá trị thực tế.

3. Xây Dựng Cây và Học Sâu Sai Lệch (Tree Building and Learning Rate):

Một cây quyết định nhỏ được xây dựng để dự đoán residuals.

Một hệ số (learning rate) nhỏ được áp dụng cho dự đoán của cây quyết định này, giảm thiểu độ lớn của ảnh hưởng của mỗi cây đối với mô hình tổng.

4. Tính Hiệu Quả Mới và Xây Dựng Cây Tiếp Theo:

Tính hiệu quả mới được tính bằng cách lấy sự chênh lệch giữa giá trị thực tế và dự đoán của mô hình hiện tại.

Cây tiếp theo được xây dựng để dự đoán tính hiệu quả mới.

Mỗi cây được xây dựng nhằm giảm độ lớn của tính hiệu quả, làm cho mô hình tổng tiến triển đến việc mô phỏng dữ liệu thực tế hơn.

5. Tổng Hợp Nhiều Cây:

Quá trình xây dựng cây và tính hiệu quả được lặp lại nhiều lần.

Kết quả cuối cùng là tổng của tất cả các cây, mỗi cây có một ảnh hưởng nhỏ dựa trên hệ số (learning rate) được áp dụng.

6. Quản lý Overfitting:

Có một số kỹ thuật được sử dụng để quản lý overfitting trong Gradient Boosting, bao gồm cắt tỉa cây (pruning), sử dụng số lượng cây thích hợp (n\_estimators), và thiết lập độ sâu tối đa cho cây (max\_depth).

7. Tham Số Quan trọng:

Các tham số quan trọng bao gồm learning rate, số lượng cây (n\_estimators), độ sâu tối đa của cây (max\_depth), và tham số liên quan đến chia nhánh cây.

Gradient Boosting thường mang lại hiệu suất cao trên nhiều bài toán regression và classification, nhưng đòi hỏi sự điều chỉnh tham số kỹ lưỡng và có thể tốn nhiều thời gian huấn luyện so với một số mô hình khác.

#### XGBRegressor

XGBoost (eXtreme Gradient Boosting) là một thuật toán máy học hàng đầu trong các cuộc thi dữ liệu và nhiều ứng dụng thực tế do khả năng mạnh mẽ của nó trong việc xử lý các vấn đề hồi quy và phân loại. XGBRegressor là một lớp mô hình trong thư viện XGBoost được sử dụng cho các bài toán hồi quy.

XGBRegressor thuộc loại ensemble learning, sử dụng nhiều cây quyết định để tạo ra một mô hình học máy mạnh mẽ.

Sử dụng cây quyết định làm cơ sở để xây dựng mô hình. Mỗi cây được xây dựng tuần tự và tập trung vào việc sửa lỗi của cây trước đó.

Dựa trên kỹ thuật Gradient Boosting, mỗi cây được xây dựng để giảm sai số (lỗi) của mô hình trước đó. Các cây sau đó cố gắng sửa lỗi của tổng của các cây trước đó.

XGBoost sử dụng một hàm mất mát cụ thể, thường là hàm bình phương sai số (Squared Error Loss) cho bài toán hồi quy.

Cung cấp các tham số chống overfitting như max\_depth, min\_child\_weight, gamma để kiểm soát độ sâu của cây và trọng số tối thiểu của mỗi lá.

XGBRegressor có một tham số quan trọng là learning\_rate điều chỉnh độ lớn của bước di chuyển trong hướng của gradient. Việc giảm learning\_rate có thể giúp cải thiện tính ổn định và chống overfitting.

Hỗ trợ subsampling (chọn một phần dữ liệu ngẫu nhiên cho mỗi cây) và column sampling (chọn một phần ngẫu nhiên của đặc trưng cho mỗi cây).

Có khả năng tự động xử lý dữ liệu hạng mục mà không cần phải chuyển đổi chúng thành số.

Hỗ trợ tính toán song song và phân phối, giúp tăng tốc quá trình đào tạo trên các tập dữ liệu lớn.

Cung cấp tính năng early stopping để dừng quá trình đào tạo khi không có sự cải thiện đủ lớn trên tập kiểm tra.

Cung cấp tính năng tích hợp để gán trọng số cho các đặc trưng, giúp điều chỉnh ảnh hưởng của từng đặc trưng đối với mô hình.

XGBRegressor là một trong những mô hình hiệu quả và linh hoạt, thường được ưa chuộng trong các bài toán hồi quy và các cuộc thi dữ liệu.

### **Quy trình huấn luyện**

## 

# 5. Results and Discussion

- Bài báo cáo này tập trung vào việc áp dụng các mô hình học máy hồi quy, bao gồm GradientBoostingRegressor, RandomForestRegressor, và XGBRegressor, để dự đoán giá trị của một biến mục tiêu trong bộ dữ liệu có 17 trường. Dựa trên kết quả trên Kaggle, nhóm đã lựa chọn 8 trường quan trọng nhất để xây dựng mô hình, là 'store\_sqft', 'florist', 'total\_children', 'avg\_cars\_at home(approx).1', 'coffee\_bar', 'num\_children\_at\_home', 'video\_store', 'salad\_bar'.

**Hiệu Suất của Các Mô Hình:**

* Kết quả cho thấy mô hình GradientBoostingRegressor đạt được RMSLE thấp nhất với giá trị 0.29985.
* RandomForestRegressor đạt được RMSLE thấp thứ 2 với giá trị 0.29773.
* XGBRegressor có RMSLE cao nhất với giá trị 0.29294.

Ưu và Nhược Điểm của Các Mô Hình:

* Gradient Boosting cho thấy hiệu suất tốt, nhưng có thể đòi hỏi thời gian đào tạo lâu hơn.
* RandomForest có kết quả tốt và thời gian đào tạo thấp hơn so với Gradient Boosting.
* XGBoost, mặc dù có RMSLE cao nhất, nhưng cũng có thể cung cấp sự linh hoạt và hiệu suất tốt trên các tập dữ liệu lớn.

# 6. Conclusion and Perspectives

**Lựa Chọn Đặc Trưng Quan Trọng:**

Việc lựa chọn 8 đặc trưng quan trọng nhất có thể giúp cải thiện tốc độ đào tạo và giảm độ phức tạp của mô hình mà vẫn giữ được hiệu suất tốt.

**Tối Ưu Hóa và Điều Chỉnh Thêm:**

Có thể tiếp tục tối ưu hóa các tham số của các mô hình và thử nghiệm các kỹ thuật điều chỉnh thêm để cải thiện hiệu suất.

**Tổng Kết và Hướng Phát Triển:**

Báo cáo cung cấp một cái nhìn tổng quan về việc sử dụng các mô hình học máy hồi quy trong bối cảnh cụ thể và đề xuất hướng phát triển tiếp theo để nâng cao kết quả dự đoán.

Tổng cộng, thông qua Báo cáo này, chúng ta đã có cái nhìn sâu sắc về việc triển khai và so sánh hiệu suất giữa các mô hình học máy hồi quy khác nhau. Kết quả và nhận xét từ bài báo cáo này có thể hỗ trợ quyết định về việc lựa chọn mô hình phù hợp trong các tình huống cụ thể.

Lời cảm ơn: Chúng em xin cảm ơn sự hỗ trợ và hướng dẫn từ cố vấn của chúng tôi trong suốt dự án này. Ngoài ra, chúng tôi muốn bày tỏ lòng biết ơn đến các tác giả của các thuật toán và thư viện được sử dụng trong nghiên cứu của chúng tôi, cũng như các nguồn của bộ dữ liệu được sử dụng.

​

# Appendix A. Project Plan management

Here is an example of Project plan, you can modify and add relevant information according to your project.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Task Name | Priority | Owner | Start date | End date | Status | Issues |
| Find documents | High | Student A | …… | …… | In progress | …… |
| Review related papers | Medium | Student A | …… | …… |  |  |
| Review and analyze public dataset | Low | Student B | …… | …… |  | Not found appropriate dataset |
| Collect and label data | High | Student B | …… | …… | …… | …… |
| Evaluate potential method | Medium | Student B | …… | …… | …… | …… |
| Experiment | Low |  | …… | …… | …… | …… |
| Compare results | Medium | Student B | …… | …… | Finished | Bad performance |
| Writing appendix | Low | Student B | …… | …… | Pending | …… |
| Future works | High | …… | …… | …… | …… | …… |

# Appendix B. Source code & Data

<https://www.kaggle.com/competitions/se355o11pmcl-08-media-campaign-cost-regression/overview>

<https://xgboost.readthedocs.io/en/stable/>

<https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why#:~:text=Random%20forest%20is%20an%20ensemble,features%20is%20selected%20for%20splitting>.

<https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machine-learning-YWOZrN7vZQ0>