

ỨNG DỤNG MACHINE LEARNING TRONG DỰ ĐOÁN TƯƠNG TÁC NGƯỜI DÙNG VỚI QUẢNG CÁO TRÊN BỘ DỮ LIỆU SOCIAL_NETWORK_ADS.CSV

Ngô Thị Loan
Trường Đại học Tiền Giang

Tóm tắt: Trong bối cảnh mạng xã hội trở thành kênh quảng bá phổ biến, việc hiểu và dự đoán hành vi tương tác của người dùng đóng vai trò quan trọng trong tối ưu hóa chiến dịch tiếp thị. Nghiên cứu này ứng dụng các thuật toán Machine Learning để phân tích và dự đoán khả năng người dùng tương tác với quảng cáo dựa trên bộ dữ liệu social_network_ads.csv. Thông qua việc trực quan hóa dữ liệu, đánh giá ảnh hưởng của các yếu tố như độ tuổi, giới tính, thu nhập, nghiên cứu xây dựng mô hình phân loại nhằm dự đoán hành vi người dùng. Kết quả cho thấy các mô hình như Random Forest, KNN và SVM có thể mang lại độ chính xác cao trong dự đoán, góp phần hỗ trợ doanh nghiệp nâng cao hiệu quả quảng cáo và tối ưu chi phí.

Từ khóa: Random Forest, KNN, SVM, hành vi người dùng, mạng xã hội.

APPLICATION OF MACHINE LEARNING IN PREDICTING USER INTERACTION WITH ADVERTISEMENTS ON THE SOCIAL_NETWORK_ADS.CSV DATASET

Ngo Thi Loan
Tien Giang University

Abstract: In the context of social networks becoming a popular advertising channel, understanding and predicting user interaction behavior plays a crucial role in optimizing marketing campaigns. This study applies Machine Learning algorithms to analyze and predict users' likelihood of interacting with advertisements based on the social_network_ads.csv dataset. By visualizing data and evaluating the impact of factors such as age, gender, and income, the research builds classification models to predict user behavior. The results show that models such as Random Forest, KNN, and SVM can achieve high accuracy in prediction, supporting businesses in improving advertising effectiveness and optimizing costs.

Keywords: Random Forest, KNN, SVM, user behavior, social networks, machine learning.

Nhận bài: 16/04/2025

Phản biện: 20/05/2025

Duyệt đăng: 23/05/2025

I. ĐẶT VẤN ĐỀ

Sự phát triển mạnh mẽ của mạng xã hội trong thời đại số đã làm thay đổi cách người dùng tiếp cận và phản hồi với các nội dung quảng cáo. Tuy nhiên, hành vi người dùng trên các nền tảng số lại ngày càng phức tạp và khó đoán định, đặt ra thách thức lớn cho các nhà tiếp thị trong việc xây dựng chiến lược phù hợp.

Với sự hỗ trợ của các thuật toán Machine Learning, việc phân tích và dự đoán hành vi người dùng không chỉ trở nên khả thi mà còn mang lại hiệu quả cao trong thực tiễn. Dữ liệu hành vi, nếu được xử lý đúng cách, sẽ giúp doanh nghiệp hiểu rõ khách hàng mục tiêu, cá nhân hóa nội dung quảng cáo và tối ưu hóa hiệu suất chiến dịch.

Nghiên cứu này lựa chọn bộ dữ liệu social_network_ads.csv – một tập dữ liệu mô phỏng hành vi người dùng trên nền tảng mạng xã hội – để thực hiện phân tích và xây dựng mô hình dự đoán khả năng tương tác với quảng cáo. Các thuật toán học máy như Random Forest, KNN, và SVM được áp dụng và so sánh nhằm xác định mô hình có độ chính xác cao nhất. Kết quả nghiên cứu không chỉ đóng góp vào lĩnh vực marketing số mà còn minh chứng hiệu quả của Machine Learning trong việc hỗ trợ ra quyết định dựa trên dữ liệu.

II. NỘI DUNG NGHIÊN CỨU

2.1. Cơ sở lý thuyết

2.1.1 Khái quát các mô hình

a. K-Nearest Neighbors (KNN)

Cơ chế: Dựa trên khoảng cách (thường là Euclidean) để tìm k điểm gần nhất và phân loại theo nhãn chiếm đa số.

Ưu điểm: Dễ hiểu, không cần huấn luyện phức tạp.

Nhược điểm: Chậm với dữ liệu lớn, kém hiệu quả với dữ liệu nhiễu hoặc nhiều chiều.

Ứng dụng: Phù hợp khi dữ liệu vừa phải và đặc trưng rõ.

b. Support Vector Machine (SVM)

Cơ chế: Tìm siêu phẳng phân tách tối ưu giữa hai lớp, dùng kernel để xử lý dữ liệu phi tuyến.

Ưu điểm: Phân loại tốt, đặc biệt với ranh giới rõ.

Nhược điểm: Cần chọn kernel phù hợp, tốn tài nguyên với dữ liệu lớn.

Ứng dụng: Phân loại hành vi có ranh giới rõ ràng.

c. Random Forest

Cơ chế: Tạo nhiều cây quyết định từ dữ liệu và đặc trưng ngẫu nhiên, kết hợp dự đoán bằng bỏ phiếu hoặc trung bình.

Ưu điểm: Chống overfitting, xử lý dữ liệu phức tạp, đánh giá được đặc trưng quan trọng.

Nhược điểm: Khó trực quan, dự đoán chậm hơn cây đơn.

Ứng dụng: Dự đoán hành vi với dữ liệu lớn và phức tạp.

2.1.2 Giới thiệu bộ dữ liệu

Bộ dữ liệu `social_network_ads.csv` (website <https://www.kaggle.com>) là tập dữ liệu mô phỏng hành vi người dùng trên mạng xã hội liên quan đến quảng cáo trực tuyến. Bộ dữ liệu gồm các thông tin cơ bản về khách hàng như:

- Tuổi (Age): độ tuổi của người dùng.
- Giới tính (Gender): nam hoặc nữ.
- Thu nhập ước tính (EstimatedSalary): thu nhập hàng năm của người dùng.
- Trang thái mua hàng hay không (Purchased): nhãn phân loại, thể hiện người dùng có tương tác mua sản phẩm sau khi xem quảng cáo hay không.

Bộ dữ liệu này được sử dụng phổ biến trong các bài toán phân tích hành vi người dùng và dự đoán tương tác với quảng cáo trên mạng xã hội. Nó có cấu trúc rõ ràng, kích thước vừa phải, thuận tiện cho việc thực hiện các kỹ thuật trực quan hóa, phân tích và áp dụng các thuật toán Machine Learning để xây dựng mô hình dự đoán hành vi.

2.2 Các nghiên cứu liên quan

Nghiên cứu này đề xuất thuật toán ODW-ELM để dự đoán hành vi người dùng trong mạng xã hội trực tuyến. Thuật toán cải thiện độ chính xác dự đoán bằng cách sử dụng trọng số cực trị trong quá trình huấn luyện, giúp nhận diện các mẫu hành vi quan trọng hơn. ACM Digital Library (*Xiong Luo và các cộng sự, 2018*); Nghiên cứu này khám phá việc kết hợp dữ liệu văn bản và hình ảnh để dự đoán mức độ tương tác của người dùng trên mạng xã hội. Nghiên cứu sử dụng các mô hình học máy như mạng nơ-ron văn bản (NN), mạng nơ-ron tích chập (CNN), Word2Vec và cây quyết định để phân tích hơn 350.000 bài đăng trên Facebook. Kết quả cho thấy việc kết hợp dữ liệu văn bản và hình ảnh mang lại hiệu quả dự đoán cao nhất (*Chad Crowe, 2024*); Nghiên cứu này sử dụng mạng nơ-ron truyền thẳng (*Feedforward Neural Network - FNN*) để dự đoán hành vi người dùng trong mạng xã hội, dựa trên các đặc trưng như nhân khẩu học, tương tác lịch sử và thuộc tính nội dung. Mô hình FNN đạt độ chính xác 87,5%, vượt trội so với các thuật toán khác trong việc phân loại hành động của người dùng như thích hoặc chia sẻ nội dung (*S.Pandikumar và các cộng sự, 2024*); Bài báo nhằm dự đoán khả năng xuất hiện thông tin giả và hành vi người dùng trên các mạng xã hội phổ

biến như Facebook, Twitter và LinkedIn. Nghiên cứu tập trung thu thập dữ liệu cá nhân từ người dùng để xây dựng hồ sơ, từ đó xác định tính cách và thông tin gốc, giúp chính phủ phát hiện người phát tán tin giả. Ứng dụng các kỹ thuật máy học hiện đại nhằm kiểm tra tính độc đáo của từng cá nhân trong tương tác xã hội. Với sự gia tăng sử dụng mạng xã hội, các công ty ngày càng cần dự báo hành vi người dùng như lượt thích, chia sẻ, bình luận để cải thiện trải nghiệm. Nghiên cứu này khám phá vai trò của máy học trong việc dự đoán hành vi người dùng trên nền tảng mạng xã hội. Nghiên cứu này đề xuất mô hình dự đoán hoạt động người dùng trên mạng xã hội, dựa trên các thuật toán học máy như Random Forest, SVM và mạng nơ-ron. Mô hình sử dụng thông tin về hồ sơ người dùng, mối quan hệ trong mạng xã hội và hành vi trước đó để dự đoán các hành động tương lai của người dùng (*Tamanna Verma và các cộng sự, 2023*); Nghiên cứu này phân tích tác động của marketing trên mạng xã hội và ứng dụng học máy trong quảng cáo kỹ thuật số, đặc biệt là trên Instagram. Nghiên cứu sử dụng các thuật toán hồi quy tuyến tính và Random Forest để dự đoán phạm vi tiếp cận của các bài đăng trên Instagram Reels, nhằm cung cấp mô hình dự đoán mạnh mẽ cho các quyết định chiến lược của doanh nghiệp (*R.Santhiya và Dr.K. Kannan, 2024*).

2.3 Phân tích kết quả đánh giá

2.3.1 Phân tích trạng thái bộ dữ liệu `Social_Network_Ads.csv` (mua hàng hay không mua hàng)

Dữ liệu tổng hợp số lượng người dùng mua và không mua theo từng độ tuổi đã được phân tích đầy đủ. Tất cả các giá trị trong dữ liệu đều đầy đủ, không có giá trị bị thiếu. Bây giờ, hãy phân tích số lượng mua và không mua theo từng độ tuổi và giới tính. Cụ thể như sau:

+ Phân tích số lượng mua hay không mua theo từng độ tuổi: Qua khảo sát cho thấy nhóm tuổi có tỷ lệ mua hàng cao nhất là 53 tuổi (100%) và nhóm tuổi có tỷ lệ không mua hàng từ 18 tuổi đến 26 tuổi (100%).

+ Phân tích số lượng mua hay không mua theo giới tính: Qua khảo sát nữ giới có tỷ lệ mua hàng là 37.75% và không mua là 62.25%, nam giới có tỷ lệ mua hàng là 33.67% và không mua là 66.33%. Như vậy, phụ nữ có xu hướng mua hàng cao hơn so với nam giới, tuy nhiên sự chênh lệch rất ít.

2.3.2 Đánh giá kết quả dự đoán dựa trên các mô hình học máy

Áp dụng một số mô hình học máy để đánh giá và dự đoán khả năng mua hàng dựa trên bộ dữ liệu. Quy trình bao gồm các bước cơ bản sau:

1. Tiền xử lý dữ liệu.
2. Chia tập dữ liệu thành tập huấn luyện, tập

kiểm tra (chuẩn hóa dữ liệu).

3. Xây dựng và huấn luyện mô hình.

4. Đánh giá hiệu suất mô hình.

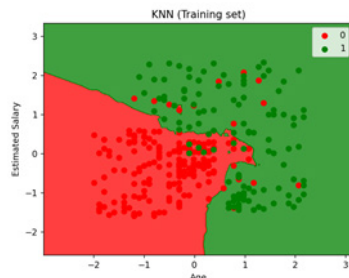
Kết quả đánh giá dự đoán các mô hình học máy như sau:

Bảng 1. Kết quả đánh giá dự đoán các mô hình học máy trên bộ dữ liệu Social_Network_Ads.csv

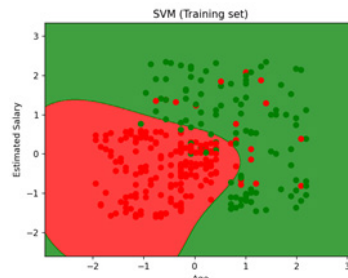
STT	Mô hình học máy	Kết quả dự đoán chính xác (Accuracy)	Ma trận sai số (Confusion Matrix) trong 100 mẫu thử
1	KNN (K-Nearest Neighbors)	93%	$\begin{bmatrix} 64 & 4 \\ 3 & 29 \end{bmatrix}$ Dự đoán đúng: $64+29=93$ Sai sót: $4+3=7$
2	SVM (Support Vector Machine)	93%	$\begin{bmatrix} 57 & 1 \\ 6 & 36 \end{bmatrix}$ Dự đoán đúng: $57+36=93$ Sai sót: $1+6=7$
3	Random Forest	91%	$\begin{bmatrix} 63 & 5 \\ 4 & 28 \end{bmatrix}$ Dự đoán đúng: $63+28=91$ Sai sót: $5+4=9$

Các mô hình huấn luyện chia bộ dữ liệu theo tỷ lệ 25% cho tập test và 75% chọn tập train. Kết quả chuẩn đoán chính xác của các mô hình là trên

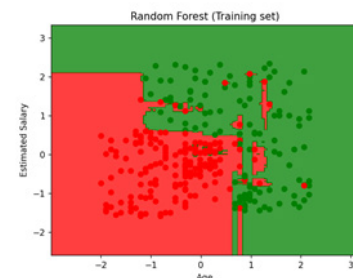
90% với các kết quả minh hoạt cho các mô hình học máy trên tập train như sau:



(a) KNN



(b) SVM



(c) Random Forest

Hình 3. Kết quả training của các mô hình học máy

Hiệu suất dự đoán lần lượt của các mô hình đều đều đạt ở mức có độ chính xác cao: mô hình KNN và SVM có độ chính xác cao nhất 93%, trong khi Random Forest đạt 91%. Tuy nhiên, KNN và SVM có sự phân bố lỗi khác nhau trong ma trận sai số: KNN có số lỗi dự đoán dương tính giả (False Positives) là 4, thấp hơn so với SVM (6). Ngược lại, SVM có số lỗi âm tính giả (False Negatives) thấp hơn (1 so với 3 của KNN).

So sánh ma trận sai số:

+ KNN: Ưu tiên hơn khi giảm thiểu lỗi dương tính giả.

+ SVM: Thích hợp hơn khi cần giảm thiểu lỗi âm tính giả.

+ Random Forest: Có độ chính xác thấp nhất

và cũng không vượt trội trong việc tối ưu hóa các loại lỗi.

Vì vậy, mô hình học máy KNN và SVM đều là lựa chọn tốt cho bộ dữ liệu này, tùy thuộc vào việc ưu tiên loại lỗi nào quan trọng hơn (dương tính giả hay âm tính giả). Random Forest mặc dù có độ chính xác thấp hơn, nhưng vẫn là một mô hình mạnh mẽ nếu cần độ ổn định hoặc dễ giải thích.

III. KẾT LUẬN

Quảng cáo trên mạng xã hội ngày càng đóng vai trò quan trọng trong chiến lược marketing của các doanh nghiệp nhằm mục đích là khách hàng đưa ra quyết định mua hàng hay không mua hàng. Việc hiểu rõ hành vi người dùng giúp tối ưu hóa quảng cáo, tăng hiệu quả chiến dịch và tiết kiệm

chi phí. Bộ dữ liệu `Social_Network_Ads.csv` là một nguồn tài nguyên quan trọng để phân tích hành vi của người dùng dựa trên các yếu tố cá nhân và xã hội.

Ứng dụng thuật toán học máy: KNN, SVM và Random Forest để dự đoán hành vi mua hàng của người dùng bằng cách chia bộ dữ liệu thành tập huấn luyện (train) và kiểm tra (test) để đánh giá hiệu suất mô hình với độ chính xác trên 90% của cả 03 mô hình. Trong đó tùy vào tính chất yêu cầu lựa chọn mô hình huấn luyện phù hợp với các đặc tính từng mô hình.

Hướng phát triển của nghiên cứu trong tương

lai là mở rộng quy mô và tính đa dạng của bộ dữ liệu, bổ sung thêm nhiều đặc trưng phản ánh hành vi người dùng như thời gian truy cập, tần suất tương tác, loại thiết bị sử dụng... Bên cạnh đó, việc ứng dụng các mô hình học sâu (Deep Learning) như MLP, RNN hoặc các kiến trúc hiện đại hơn như Transformer có thể giúp cải thiện độ chính xác dự đoán. Ngoài ra, cần thực hiện tối ưu hóa siêu tham số, đánh giá mô hình với dữ liệu thực tế, từ đó xây dựng hệ thống hỗ trợ ra quyết định quảng cáo hiệu quả hơn, giúp doanh nghiệp cá nhân hóa chiến dịch và tiết kiệm chi phí quảng bá trên mạng xã hội.

TÀI LIỆU THAM KHẢO

- [1] Xiong, L., Weiping, W., Changwei, J. and Yang, X. (2018). User behavior prediction in social networks using weighted extreme learning machine with distribution optimization, *Future Generation Computer Systems, Volume 93*.
- [2]. Crowe C. (2018). Predicting user interaction on social media using machine learning (Master's thesis, University of Nebraska at Omaha). *DigitalCommons@UNO*.
- [3]. Pandikumar, S., Menaka, C., & Sevugapandi, N. (2024). Behavior prediction in social networks using feedforward neural network algorithm. In *Innovations and trends in modern computer science technology – Overview, challenges and applications*, (pp. 1–8). <https://doi.org/10.48001/978-81-980647-5-2-1>.
- [4]. Verma, T., Gupta, R. K., Kumari, D., & Soni, K. (2023). User activity prediction on social network using machine learning. In *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 11(5). <https://doi.org/10.22214/ijraset.2023.50311>.
- [5]. Santhiya, R., and Kannan, K. (2024). The impact of social media marketing and machine learning applications in digital advertising. *Journal of Philanthropy and Marketing*.