Phân tích phim để dự đoán khả năng thư ơ ng mại của chúng đối với nhà sản xuất

Devendra Swami
Đại học Nam California Los Angeles, California
dswami@usc.edu

Yash Phogat
Đại học Nam California Los Angeles, California
phogat@usc.edu

Aadiraj Batlaw
Đại học California Berkeley, California batlaw33375@berkeley.edu

Ashwin Goyal
Đại học Nam California Los Angeles, California
ashwingo@usc.edu

TÓM TẮT

Khi bộ phim ra mắt, một hình thức suy đoán chính liên quan đến thành công tư ơ ng đối của bộ phim. Tính tư ơ ng đối này đặc biệt liên quan đến ngân sách ban đầu của bộ phim – vì nhiều lần " phim bom tấn kinh phí lớn" đã đạt đư ợc thành công đặc biệt cũng như thất bại thảm hại. Vậy làm thế nào để dự đoán đư ợc thành công của một bộ phim sắp ra mắt? Trong bài báo này, chúng tôi đã khám phá một mảng dữ liệu phim lớn trong nỗ lực phát triển một mô hình có thể dự đoán lợi nhuận dự kiến của một bộ phim sắp ra mắt. Cách tiếp cận để phát triển này như sau: Đầu tiên, chúng tôi bắt đầu với tập dữ liệu MovieLens [2] có các thuộc tính phim chung cùng với thẻ bộ gen cho mỗi phim. Thẻ bộ gen cung cấp thông tin chi tiết về những đặc điểm cụ thể nào của bộ phim là nổi bật nhất. Sau đó, chúng tôi đã đư a vào các tính năng bổ sung liên quan đến nội dung phim, dàn diễn viên/đoàn làm phim, nhận thức của khán giả, ngân sách và thu nhập từ các trang web TMDB, IMDB và Metacritic. Tiếp theo, chúng tôi đã thực hiện phân tích dữ liệu khám phá và thiết kế một loạt các tính năng mới để nắm bắt thông tin lịch sử cho các tính năng có sẵn. Sau đó, chúng tôi đã sử dụng phân tích giá trị kỳ dị (SVD) để giảm chiều của các tính năng có chiều cao (ví dụ: thẻ bộ gen). Cuối cùng, chúng tôi đã xây dựng một Bộ phân loại Rừng ngẫu nhiên và thực hiện điều chỉnh siêu tham số để tối ư u hóa độ chính xác của mô hình. Ứng dụng trong tư ơ ng lai của mô hình của chúng tôi có thể đư ợc thấy trong ngành công nghiệp phim ảnh, cho phép các công ty sản xuất dự đoán tốt hơ n lợi nhuận dự kiến của các dự án dựa trên phác thảo về quy trình sản xuất của họ, do đó cho phép họ sửa đổi kế hoạch để đạt đư ợc lợi nhuận tối ư u.

1 GIỚI THIỆU

Ngành công nghiệp điện ảnh là một ngành nổi bật ở mọi khía cạnh. Đây là một thế giới riêng. Đối với bài báo này, động lực của chúng tôi xuất phát từ mong muốn cung cấp một mô hình dự đoán cho các nhà sản xuất để có đư ợc ý tư ởng về khả năng thư ơ ng mại của bộ phim mà họ đề xuất. Joe Swanberg đã nói vào năm 2016, "Cách duy nhất để bạn kiếm đư ợc tiền là nếu bạn đầu tư vào các bộ phim của chính mình". Trư ớc khi các nhà sản xuất phim phải hoàn tất quyết định của mình, họ phải đảm bảo rằng khoản đầu tư của họ là hợp lý và hiểu cách họ thấy đư ợc lợi nhuận từ khoản đầu tư đó - đây là nơ i mô hình của chúng tôi bư ớc vào thế giới điện ảnh.

Những đóng góp chính trong việc giải quyết vấn đề này có thể đư ợc tóm tắt như sau:

- Chúng tôi đã nghiên cứu nhiều đặc điểm có khả năng liên quan đến thành công thư ơ ng mại của một bộ phim. Khác với các nghiên cứu có sẵn khác, chúng tôi đã kết hợp nhiều đặc điểm mới lạ như công khai, ngày phát hành và dàn diễn viên & đoàn làm phim. Chúng tôi đã dành nhiều thời gian vào kỹ thuật tính năng để hiểu rõ hơ n những yếu tố nào khiến một bộ phim có lợi nhuận về mặt tài chính.

- Chúng tôi đã trích xuất 11 nhóm tính năng khác nhau và xây dựng một mô hình rừng ngẫu nhiên (RF) để dự đoán lợi tức đầu tư (ROI) cho bộ phim sẽ cao hơ n hay thấp hơ n mức trung bình. Sau khi đào tạo RF, chúng tôi đã xác định đư ợc tầm quan trọng tư ơ ng đối của từng tính năng riêng lẻ và các nhóm tính năng.

Bài báo còn lại đư ợc cấu trúc như sau. Trong Phần 2, chúng tôi mô tả phư ơ ng pháp nghiên cứu của mình. Trong Phần 3, chúng tôi trình bày những phát hiện của mình, tiếp theo là Phần 4, nơ i chúng tôi thảo luận về mối quan hệ giữa một số tính năng quan trọng và ROI. Trong Phần 5, chúng tôi xác định các mối đe dọa đối với tính hợp lệ và cuối cùng trong Phần 6, chúng tôi kết luận bài báo và đề xuất phạm vi cho công việc trong tư ơ ng lai.

2 PHƯ Ơ NG PHÁP NGHIÊN CỨU

2.1 Phát biểu vấn đề

Bài báo hiện tại nghiên cứu ứng dụng của việc sử dụng thuật toán học máy để cung cấp hiểu biết tốt hơ n về các tính năng có khả năng ảnh hư ởng đến thành công thư ơ ng mại của một bộ phim. Phát biểu vấn đề có thể đư ợc định nghĩa chính thức như sau Nhiệm vụ: Dự đoán thành công của bộ phim: Cho một bộ phim , dự đoán liệu đây có phải là một bộ phim thành công về mặt thư ơ ng mại hay không. Trong khi thực hiện nhiệm vụ này, chúng tôi cố gắng trả lời các câu hỏi nghiên cứu đư ợc đề cập dư ới đây.

- RQ1: Thuật toán rừng ngẫu nhiên có thành công như thế nào trong việc dự đoán liệu một bộ phim có thành công về mặt thư ơ ng mại xét về mặt ROI hay không?

- RQ2: Những tính năng và nhóm tính năng riêng lẻ nào đóng vai trò quan trọng nhất trong việc dự đoán ROI từ phim ảnh?

2.2 Tính năng

Trong phần này, chúng tôi thảo luận về các tính năng mà chúng tôi đã xem xét trong nghiên cứu này. Các tính năng đư ợc lựa chọn dựa trên các tính năng thư ờng đư ợc sử dụng để thực hiện phân tích ngành công nghiệp phim ảnh và các tính năng khác

Giải nhì, Khu vực Bờ Tây, Citadel Securities, tháng 10 năm 2020, CA, Hoa Kỳ

các tính năng mới có sẵn công khai và có thể dễ dàng trích xuất bằng các công cụ có thể truy cập đư ợc.

Bảng 1 tóm tắt các tính năng đư ợc xem xét trong nghiên cứu này. Chúng tôi phân loại các tính năng đã chọn thành 11 nhóm dựa trên đặc điểm của một tính năng trong nhóm. Mỗi nhóm tính năng đư ợc tóm tắt dư ới đây.

2.2.1 Nội dung. Trong nhóm này, chúng tôi xem xét is_adult, is_english, languages_count, runtime, genome và genre. Nội dung của phim rất quan trọng vì ngư ời xem phim thư ờng có sở thích riêng biệt đối với các danh mục này (ví dụ: ngư ời nói tiếng Anh thư ờng thích phim tiếng Anh), do đó nội dung phim cho phép chúng tôi thu hẹp đối tư ợng mục tiêu.

2.2.2 Quảng cáo. Một số tính năng đã đư ợc trích xuất để đo lư ờng các nỗ lực quảng cáo, như is_collection, is_homepage, is_tagline và key- words_count. Quảng cáo là một yếu tố quan trọng, vì tiếp thị và khả năng tiếp cận của bộ phim có tác động trực tiếp đến số lư ợng ngư ời nghe về bộ phim và theo đó là số lư ợng ngư ời xem bộ phim (đặc biệt là ngày nay, các bộ phim thuộc loạt phim có xu hư ớng vư ợt trội hơ n các bộ phim khác về doanh thu phòng vé).

2.2.3 Nhận thức của khán giả. Trong nhóm này, chúng tôi đã xem xét các tính năng bao gồm mức độ phổ biến, vote_average, vote_count, metacritic_score, imdb_rating, imdb_votes. Đư ơ ng nhiên, nhận thức của khán giả có tác động rất lớn đến thành công của phim. Những bộ phim có xếp hạng cao có xu hư ớng vư ợt trội hơ n những bộ phim có xếp hạng thấp hơ n. Tuy nhiên, vì thông tin này thư ờng không có sẵn cho đến sau khi bộ phim đư ợc phát hành, nên chúng tôi đã không đư a thông tin này vào mô hình dự đoán của mình.

2.2.4 Ngày phát hành. Trong nhóm này, chúng tôi đã kiểm tra các tính năng như release_month, movies_per_month, budget_fraction, expense_score.
Một số nghiên cứu đã chỉ ra rằng sự cạnh tranh mà một bộ phim gặp phải tại thời điểm phát hành có tác động lớn đến thành công về mặt tài chính của bộ phim. Nhóm tính năng này cố gắng xem xét hiệu ứng này.

2.2.5 Tài chính. Nhóm này chỉ bao gồm các tính năng, time_discounted_budget, là giá trị chiết khấu của ngân sách - đư ợc sử dụng để sản xuất phim. Thông thư ờng, những bộ phim có kinh phí lớn hơ n có xu hư ớng vư ợt trội hơ n những bộ phim có kinh phí thấp hơ n về mặt doanh thu.

2.2.6 Nhà sản xuất. Nhóm này chỉ bao gồm tính năng production_house_embedding. Tính năng này đư ợc tính toán từ hiệu suất trung bình của các bộ phim gần đây do cùng một nhà sản xuất sản xuất.
Các hãng sản xuất phim tốt và uy tín hơ n thư ờng có thể thuê đư ợc nhiều đạo diễn, biên kịch và ngôi sao nổi tiếng hơ n, ngoài việc có ngân sách lớn hơ n, đây cũng là những yếu tố thư ờng dẫn đến thành công về mặt doanh thu phòng vé.

2.2.7 Biên kịch. Nhóm này chỉ bao gồm tính năng writ-ers_embedding, đư ợc tính toán từ hiệu suất trung bình của các bộ phim gần đây do cùng một biên kịch viết. Thông thư ờng, các biên kịch đã viết cho những bộ phim thành công có xu hư ớng viết thành công hơ n phim ảnh.

2.2.8 Đạo diễn. Nhóm này chỉ bao gồm các đạo diễn đặc trư ng_embedding, đư ợc tính toán từ hiệu suất trung bình của các bộ phim gần đây do cùng một đạo diễn thực hiện. Thông thư ờng, các đạo diễn đã chỉ đạo các bộ phim thành công có xu hư ớng chỉ đạo các bộ phim thành công hơ n phim ảnh.

2.2.9 Nhà sản xuất. Nhóm này chỉ bao gồm các tính năng producers_embedding, đư ợc tính toán từ hiệu suất trung bình của các bộ phim gần đây do cùng một nhà sản xuất thực hiện. Thông thư ờng, các nhà sản xuất đã sản xuất những bộ phim thành công có xu hư ớng sản xuất những bộ phim thành công hơ n.

2.2.10 Diễn viên chính. Nhóm này chỉ bao gồm tính năng main_cast_embedding, đư ợc tính toán từ hiệu suất trung bình của các bộ phim gần đây mà các thành viên từ dàn diễn viên chính tham gia
Nhiều nghiên cứu đã chỉ ra rằng ngôi sao của bộ phim có ảnh hư ởng lớn đến sự thành công của bộ phim.

2.2.11 Nhân viên hỗ trợ. Nhóm này bao gồm các tính năng fe-male_count, male_count và crew_length. Một số nghiên cứu chỉ ra rằng giới tính và quy mô của dàn diễn viên trong một bộ phim có thể ảnh hư ởng đến thành công.

2.3 Thu thập dữ liệu

Chúng tôi đã lấy đư ợc hầu hết dữ liệu từ tập dữ liệu Kaggle "The Movies Dataset",[3] cung cấp cho chúng tôi dữ liệu về dàn diễn viên, đoàn làm phim và siêu dữ liệu chung trên một tập hợp con lớn các bộ phim. Trong một tập dữ liệu do cộng đồng cuộc thi Data Open cung cấp, chúng tôi đã lấy đư ợc các thẻ bộ gen, chúng tôi đã hợp nhất chúng với tập dữ liệu siêu dữ liệu của mình. Chúng tôi đã lấy thêm các tính năng bằng cách thực hiện các lệnh gọi API đến TMDB (The Movie Database), cũng như thu thập điểm Metacritic từ "IMDB: All US Phim đã phát hành: 1972-2016". Ban đầu, tập dữ liệu hợp nhất của chúng tôi bao gồm hơ n 13 nghìn hàng một chút vì chúng tôi chỉ có thông tin bộ gen cho nhiều bộ phim này. Sau đó, chúng tôi quyết định tập trung vào các bộ phim sau cuộc cách mạng màu (tức là sau năm 1965) và loại bỏ các hàng có giá trị ngân sách hoặc giá trị doanh thu đư ợc đặt thành 0. (Tuy nhiên, chúng tôi đã cân nhắc trư ớc để nội suy các lỗi này, vì ngân sách và doanh thu là một yếu tố quan trọng trong việc xác định tỷ lệ hoàn vốn, nên chúng tôi kết luận rằng bất kỳ phư ơ ng pháp nội suy nào cũng sẽ làm sai lệch mô hình của chúng tôi phần lớn). Điều này đư a tổng số hàng của chúng tôi lên 5.426.

2.4 Thuật toán học máy Lúc đầu chúng tôi nghĩ đến việc sử dụng hồi quy làm thuật toán học máy của mình do thực tế là các giá trị ROI là liên tục. Tuy nhiên, trong nhiều trư ờng hợp, dự đoán của chúng tôi từ hồi quy không chính xác. Hơ n nữa, thay vì quan tâm đến việc biết các giá trị chính xác, các nhà sản xuất phim có thể có xu hư ớng muốn biết liệu bộ phim của họ có khả năng hoạt động tốt hay không. Do đó, chúng tôi đã quyết định dự đoán trên hoặc dư ới ROI trung bình (trung bình đư ợc tính trên dữ liệu đào tạo) thay vì các giá trị ROI chính xác. Theo cách này, chúng tôi đã chuyển đổi vấn đề hồi quy khó học ban đầu của mình thành một nhiệm vụ phân loại nhị phân dễ dàng.

Chúng tôi đã triển khai thuật toán rừng ngẫu nhiên (RF) để thực hiện nhiệm vụ phân loại của mình vì đây là một trong những thuật toán phân loại phi tuyến tính thành công nhất. Thuật toán học máy. Nó xem xét nhiều cây quyết định đư ợc đào tạo trên mẫu ngẫu nhiên của dữ liệu đào tạo. Ngoài ra, để phân chia các nút, nó chọn ngẫu nhiên một tập hợp con các tính năng. Cuối cùng, quyết định đư ợc thực hiện bằng cách lấy trung bình các dự đoán từ mỗi cây quyết định. Do đó, thuật toán rừng ngẫu nhiên ít có khả năng bị quá khớp. Hơ n nữa, nó có thể lấy các biến số hoặc biến danh mục làm đầu vào và thậm chí không yêu cầu mở rộng tính năng. Vì những lý do này, chúng tôi đã chọn thuật toán rừng ngẫu nhiên để thực hiện nhiệm vụ phân loại. Tập dữ liệu đư ợc chia thành khoảng 80% đào tạo

Bảng 1: Các tính năng của phim có khả năng ảnh hư ởng đến thành công thư ơ ng mại của một bộ phim

#	Nhóm	Tính năng	Sự miêu tả
1	Nội dung	là_ngư ời_lớn	Bộ phim này có chỉ phù hợp với ngư ời lớn hay không.
		là_tiếng_anh	Ngôn ngữ chính đư ợc sử dụng trong phim có phải là tiếng Anh hay không.
		ngôn	Tổng số ngôn ngữ đư ợc sử dụng trong phim.
		ngữ_số_lư ợng_phim_runtime	Tổng độ dài của bộ phim, tính bằng phút.
		thẻ_genome [6]	Các tính năng đư ợc trích xuất từ nội dung phim.
		thể loại phim	Những thể loại khác nhau mà bộ phim có thể đư ợc xếp vào.
2	Công khai	là_bộ_sư u_cung	Bộ phim có phải là một phần của bộ sư u tập hay loạt phim không.
		là trang chủ	Bộ phim có trang chủ hay không.
		là_khẩu_hiệu	Bộ phim có khẩu hiệu liên quan hay không.
		số lư ợng từ khóa	Số lư ợng từ khóa thư ờng dùng có thể gán cho phim.
3	Nhận thức của khán giả†	sự phổ biến	Điểm số phổ biến của phim do TMDB cung cấp.
		bình chọn_trung_bình	Điểm chấp thuận trung bình cho bộ phim, dựa trên tổng số phiếu bầu tại TMDB.
		số phiếu bầu	Tổng số phiếu bầu cho bộ phim tại TMDB.
		Điểm Metacritic	Điểm đánh giá trung bình của phim trên Metacritic.
		Xếp hạng imdb	Điểm đánh giá trung bình của phim trên IMDB.
		imdb_vote	Tổng số phiếu bầu cho bộ phim trên IMDB.
4	Ngày phát hành	tháng phát hành	Tháng mà bộ phim đư ợc phát hành.
		movies_per_month	Tổng số phim đư ợc phát hành trong cùng tháng với bộ phim đư ợc chỉ định.
		budget_fraction	Ngân sách của bộ phim tỷ lệ thuận với ngân sách tích lũy của tất cả các bộ phim trong tháng đó.
		điểm_chi_phí_phim	Ngân sách của bộ phim tư ơ ng ứng với ngân sách trung bình của tất cả các bộ phim trong tháng đó.
5	Tài chính	time_discounted_budget	Giá trị chiết khấu theo thời gian cho ngân sách dùng để sản xuất phim.
6	Nhà sản xuất	nhà sản xuất_nhúng_writer_nhúng	Tính toán dựa trên hiệu suất của các bộ phim gần đây do cùng một hãng sản xuất sản xuất.
7	Nhà văn		Tính toán dựa trên thành tích của những bộ phim gần đây do cùng tác giả chấp bút.
8	Giám đốc	giám đốc_nhúng	Tính toán dựa trên thành tích của những bộ phim gần đây do cùng đạo diễn chỉ đạo.
9	Nhà sản xuất	nhà sản xuất_nhúng	Tính toán dựa trên hiệu suất của các bộ phim gần đây do cùng một nhà sản xuất sản xuất.
10	Diễn viên chính	nhúng main_cast	Tính toán dựa trên thành tích của các bộ phim gần đây có sự tham gia của các thành viên trong dàn diễn viên chính.
11	Nhân viên hỗ trợ	nữ_số_lư ợng	Tổng số diễn viên nữ trong phim.
		nam_số_lư ợng	Tổng số diễn viên nam trong phim.
		chiều dài phi hành đoàn	Tổng số ngư ời trong đoàn làm phim.

†: Không đư ợc sử dụng trong mô hình dự đoán vì thông tin đó không có sẵn trư ớc khi phim đư ợc phát hành.

và 20% dữ liệu thử nghiệm với các bộ phim trư ớc năm 2011 là một phần của chư ơ ng trình đào tạo dữ liệu và từ năm 2011 trở đi bao gồm bộ thử nghiệm.

Giảm chiều: Do sự hiện diện của các đặc điểm thư a thớt có chiều cao như thẻ bộ gen và thông tin thể loại (sau một lần mã hóa nóng) trong tập dữ liệu, chúng tôi đã sử dụng giá trị kỳ dị phân tích (SVD) để giảm số chiều của chúng trong khi giữ nguyên hầu hết các phư ơ ng sai trong dữ liệu. Ngoài ra, chúng tôi cũng đã loại bỏ các tính năng có tư ơ ng quan cao khỏi tập dữ liệu của chúng tôi. Chúng tôi đã xác định đư ợc cặp tính năng có mối tư ơ ng quan cao, tức là cặp có giá trị tư ơ ng quan tuyệt đối là 0,75 hoặc cao hơ n từ tư ơ ng quan Spearman . Trong số các cặp đư ợc xác định này, các tính năng có mối quan hệ tư ơ ng hỗ thấp hơ n thông tin [4] bị loại bỏ. Nó đư ợc thực hiện để loại bỏ các tính năng không không cung cấp bất kỳ thông tin bổ sung nào và có thể giúp giảm kích thư ớc của tập dữ liệu để tăng tốc quá trình đào tạo của chúng tôi.

Tối ư u hóa siêu tham số: Để có đư ợc siêu tham số tối ư u cho mô hình RF của chúng tôi, chúng tôi đã bắt đầu bằng cách chuẩn bị một không gian tìm kiếm lư ới. Đối với n_estimators, chúng tôi đã xem xét các giá trị từ 100 đến 1000 với bư ớc nhảy là 100. Tư ơ ng tự như vậy, bội số nguyên của √ # đến #features cho max_features, giá trị từ 10 đến 100 với kích thư ớc bư ớc là 10 cho max_depth đư ợc sử dụng. Các giá trị 0,01, 0,03, và 0,05 đư ợc xem xét cho min_samples_split, và 1, 3 và 5 cho min_samples_leaf. Chúng tôi luôn đặt bootstrap thành True.

Do không gian tìm kiếm lư ới của 12600 (10 · 14 · 10 · 3 · 3) là tốn kém về mặt tính toán, chúng tôi đã thực hiện tìm kiếm ngẫu nhiên trên các mẫu đư ợc rút ra đồng đều từ nó. Chúng tôi đã sử dụng 100 lần lặp lại thông qua xác thực chéo 4 lần để xác định các siêu tham số tối ư u. Chúng tôi báo cáo các siêu tham số thu đư ợc từ quy trình này trong Bảng 2.

Bảng 2: Các tham số của mô hình RF đư ợc điều chỉnh trên tập xác thực

Các tham số	Giá trị tối ư u
n_estimators	500
max_features	14
max_depth	40
min_samples_split	0,05
min_samples_leaf	5

2.5 Đánh giá mô hình

Nhờ cách chúng tôi xây dựng nhiệm vụ phân loại của mình, tập dữ liệu của chúng tôi đư ợc cân bằng hoàn hảo với hai lớp có tầm quan trọng như nhau. Do đó , Độ chính xác ban đầu đư ợc coi là một thư ớc đo đánh giá phù hợp.
Tuy nhiên, độ chính xác phụ thuộc vào việc lựa chọn ngư ỡng để chuyển đổi xác suất dự đoán thành lớp đầu ra có khả năng xảy ra cao nhất. Khác nhau ngư ỡng có thể tạo ra điểm chính xác khác nhau cho cùng một dự đoán ML và do đó gây khó khăn cho việc so sánh hiệu suất trên nhiều thuật toán ML.

Để vư ợt qua rào cản này, chúng tôi đã chọn Diện tích dư ới Đư ờng cong Đặc tính hoạt động của Máy thu (ROC) để đánh giá hiệu suất thực hiện nhiệm vụ của chúng tôi. Chúng tôi sử dụng từ viết tắt AUC để biểu thị số liệu này. Đư ờng cong ROC là một biểu đồ giữa tỷ lệ dư ơ ng thực sự và tỷ lệ dư ơ ng tính giả thu đư ợc bằng cách triển khai các ngư ỡng khác nhau để chỉ định các giá trị dư ơ ng. Diện tích dư ới đư ờng cong này (AUC) sau đó đư ợc sử dụng để so sánh hiệu suất của một mô hình trên phân loại nhị phân của chúng tôi nhiệm vụ.

Hơ n nữa, chúng tôi đã so sánh hiệu suất của thuật toán rừng ngẫu nhiên với phư ơ ng pháp tiếp cận đư ờng cơ sở ngẫu nhiên. Trong Ngẫu nhiên đư ờng cơ sở, chúng tôi ngẫu nhiên chỉ định dư ới hoặc trên ROI trung bình cho mỗi phim trong bộ thử nghiệm. Không có sự cải thiện đáng kể nào trong hiệu suất của mô hình RF trên đư ờng cơ sở sẽ ngụ ý rằng đã chọn các tính năng không liên quan đến thành công thư ơ ng mại của bộ phim.

Một mô hình phân loại hoàn hảo sẽ mang lại giá trị AUC là 1,00, trong khi một mô hình ngẫu nhiên sẽ có giá trị khoảng 0,50. Nói chung, một giá trị AUC cao hơ n cho thấy mô hình tốt hơ n.

2.6 Phân tích tầm quan trọng của tính năng

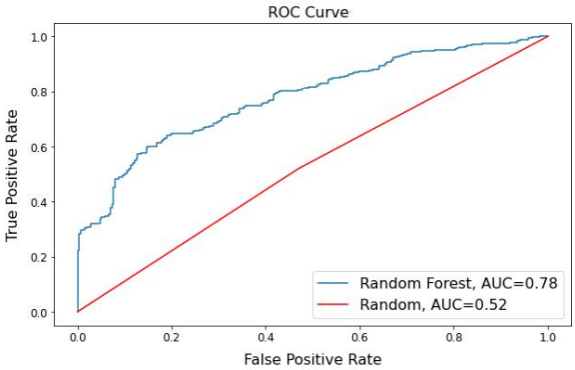
Chúng tôi đã sử dụng kỹ thuật tính năng hoán vị quan trọng đư ợc đề xuất bởi Breiman [1] để đo lư ờng tầm quan trọng của một tính năng riêng lẻ và một nhóm tính năng. Ý tư ởng chính đằng sau phư ơ ng pháp này là rằng sự hoán vị ngẫu nhiên của các tính năng quan trọng sẽ dẫn đến một sự suy giảm đáng kể về hiệu suất của mô hình. Sự suy giảm này thư ớc đo hiệu suất đư ợc gọi là Giá trị quan trọng (IV) và giá trị càng cao thì tính năng đó càng quan trọng.

Tư ơ ng tự như vậy, tầm quan trọng của nhóm đư ợc tính bằng cách hoán vị tất cả các tính năng trong một nhóm với nhau. Điều quan trọng là không ở đây rằng chúng ta có thể sử dụng an toàn kỹ thuật quan trọng của tính năng hoán vị [5] vì không có tính năng nào có mối tư ơ ng quan cao trong tập dữ liệu của chúng tôi chúng tôi đã loại bỏ những tính năng có mối tư ơ ng quan cao.

3 KẾT QUẢ

3.1 RQ1: Rừng ngẫu nhiên thành công như thế nào thuật toán trong việc dự đoán xem một bộ phim sẽ là một thành công thư ơ ng mại về mặt Lợi tức đầu tư ?

Hình 1 cho thấy các đư ờng cong ROC của các đư ờng cơ sở và ngẫu nhiên Thuật toán rừng. AUC cho thuật toán rừng ngẫu nhiên là 0,78.
Về cơ sở, Random đạt đư ợc điểm AUC mong đợi khoảng 0,500. Điểm AUC cao hơ n của mô hình RF cho thấy rằng giải pháp sử dụng kỹ thuật tính năng và máy học sẽ có khả năng giúp các nhà sản xuất phim và các nhà phân tích làm tốt hơ n ư ớc tính lợi tức đầu tư dự kiến cho các bộ phim sắp ra mắt.



Hình 1: Đư ờng cong ROC của thuật toán rừng ngẫu nhiên sử dụng các tính năng đã chọn và đư ờng cơ sở cho tập dữ liệu của chúng tôi.

3.2 RQ2: Những đặc điểm và nhóm riêng lẻ nào của các tính năng đóng vai trò quan trọng nhất trong dự đoán ROI từ phim ảnh?

Bảng 3: Tầm quan trọng của 15 tính năng cá nhân hàng đầu

Tên tính năng	Giá trị thông tin (IV)
is_collection	0,010
bộ gen_0 số	0,005
lư ợng từ khóa bộ	0,004
gen_2 bộ	0,004
gen_13	0,004
phim_mỗi_tháng số lư ợng	0,003
nam bộ gen_3	0,002
bộ gen_5 bộ	0,002
gen_12	0,002
is_homepage	0,002
thời	0,001
gian_giảm_giá_ngân_sách_female_count	0,001
bộ gen_1 bộ gen_10	0,001
	0,001
	0,001

Bảng 3 cho thấy 15 tính năng quan trọng nhất trong việc dự đoán ROI của phim. Chúng tôi cũng quan sát thấy is_collection nằm trong số tính năng quan trọng nhất. Điều này có ý nghĩa trực quan, như phần tiếp theo, các phim phụ và phim mở rộng của một vũ trụ chủ đề có xu hư ớng vư ợt trội hơ n các phim không có những đặc điểm như vậy. Chúng tôi cũng nhận thấy rằng các đặc điểm của bộ gen (mà như ngư ời ta có thể nhớ lại mô tả một bộ phim của (đặc điểm nổi bật nhất của nó) có xu hư ớng cung cấp thông tin có giá trị. key_word_count cung cấp thông tin tư ơ ng tự như tính năng bộ gen, và nó cũng có giá trị thông tin cao ngụ ý rằng đặc điểm của phim là một cái nhìn sâu sắc quan trọng để xác định sự thành công của một bộ phim. Quan sát đáng chú ý cuối cùng trong bảng này sẽ là movies_per_month, ngụ ý rằng sự cạnh tranh vào ngày phát hành lại của một bộ phim thực sự là một yếu tố chính góp phần quyết định sự thành công của một bộ phim.

Bảng 4 hiển thị 5 nhóm tính năng quan trọng nhất. Ở đây chúng tôi quan sát rằng Nội dung có giá trị thông tin cao nhất. Điều này làm cho

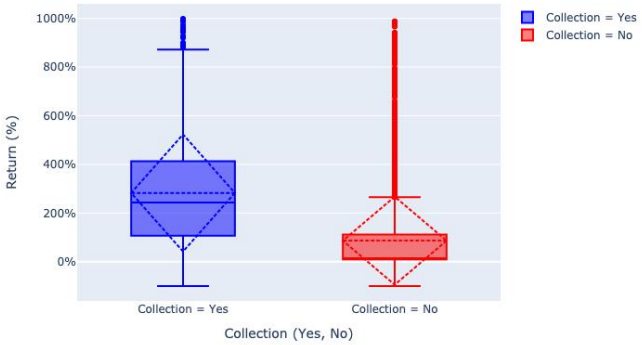
Bảng 4: Tầm quan trọng của 5 nhóm tính năng hàng đầu

Giá trị thông tin nhóm	tính năng (IV)
Nội dung	0,047
Diễn viên chính	0,031
Công khai	0,021
Nhà văn	0,017
Nhà sản xuất	0,016

nghĩa là, vì genome_tags đư ợc đư a vào như một tính năng trong danh mục này. Với Main Cast đứng thứ hai trong IV, chúng tôi khẳng định lại giả thuyết của mình rằng các ngôi sao của một bộ phim có tác động lớn đến thành công của bộ phim.

4 THẢO LUẬN

Thay vì chỉ dựa vào các giá trị tầm quan trọng của tính năng đã thảo luận ở phần trư ớc, chúng tôi cũng muốn tìm hiểu mối quan hệ nhân quả và xác định hư ớng ảnh hư ởng của các tính năng quan trọng lên giá trị ROI. Do đó, trong phần này, chúng tôi sẽ khám phá cụ thể mối quan hệ giữa các tính năng chính và lợi nhuận. Tuy nhiên, chúng tôi lư u ý rằng đây là phân tích đơ n biến có như ợc điểm là không xem xét các tư ơ ng tác đa biến.

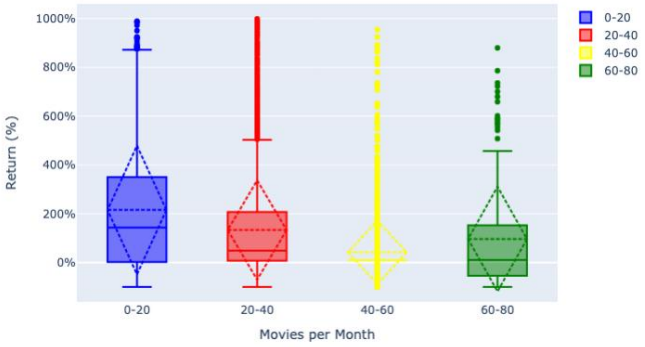


Hình 2: Biểu đồ giữa Is_collection và Return

Hình 2 cho thấy có vẻ như có sự khác biệt hợp lý về lợi tức đầu tư của những bộ phim nằm trong bộ sư u tập và những bộ phim không nằm trong bộ sư u tập. Những bộ phim nằm trong bộ sư u tập/loạt phim có xu hư ớng mang lại ROI cao hơ n những bộ phim không nằm trong bất kỳ bộ sư u tập nào.

Hình 3 cho thấy có vẻ như có sự khác biệt khá hợp lý giữa lợi tức đầu tư cho các bộ phim đư ợc phát hành với các phạm vi khác nhau về số lư ợng phim đư ợc phát hành trong tháng đó. Có vẻ như phần lớn, càng ít phim đư ợc phát hành trong tháng của ngày phát hành phim quan tâm thì tỷ lệ lợi nhuận càng cao. Điều quan trọng cần lư u ý là có những ngoại lệ lớn cho tất cả các danh mục này, báo hiệu rằng một số phim trên tất cả các danh mục đều có thành tích đặc biệt tốt bất kể số lư ợng phim đư ợc phát hành trong cùng tháng của ngày phát hành phim.

Hình 4 khám phá xem sự tồn tại của trang chủ cho một bộ phim cụ thể có liên quan đến lợi tức đầu tư cao hay không. Từ hình này, có vẻ như có mối liên hệ, như ng có vẻ như không đặc biệt quan trọng đối với trư ờng hợp đơ n biến vì chúng ta thấy các khoảng tin cậy chồng chéo.

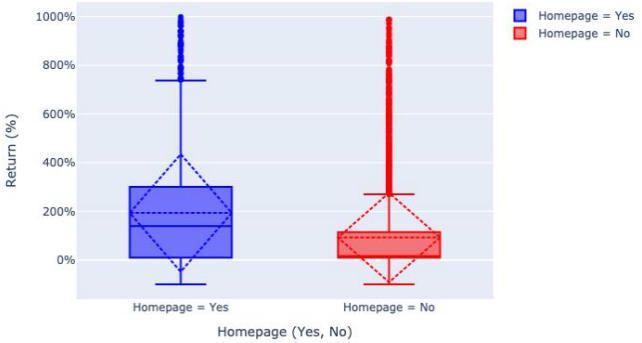


Hình 3: Biểu đồ giữa Phim mỗi tháng và Lợi nhuận

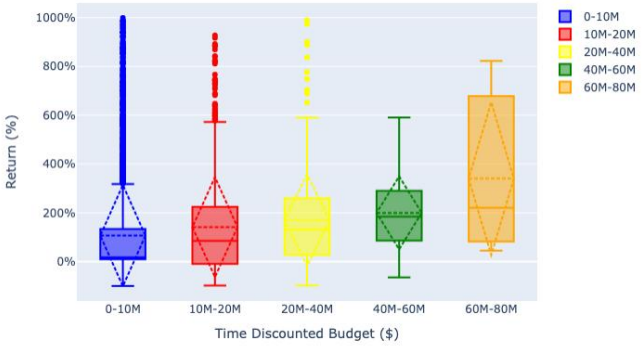
Hình 5 so sánh lợi tức đầu tư cho các loại ngân sách khác nhau. Từ đây chúng ta thấy rằng các bộ phim đư ợc sản xuất với ngân sách lớn hơ n có xu hư ớng có lợi tức đầu tư cao hơ n.

5 MỐI ĐE DỌA ĐỐI VỚI TÍ NH HỢP LỆ

Chúng tôi đã liệt kê nhiều mối đe dọa khác nhau đến tính hợp lệ và hạn chế của nghiên cứu trong phần này.



Hình 4: Biểu đồ giữa Is_homepage và Return



Hình 5: Biểu đồ giữa Ngân sách chiết khấu theo thời gian và Lợi nhuận

Độ lệch lựa chọn tính năng: Kết quả của nghiên cứu hiện tại có thể bị ảnh hư ởng bởi các tính năng mà chúng tôi đã xem xét, bị giới hạn bởi trí tư ởng tư ợng của chúng tôi . Do đó, những ngư ời khác sử dụng một bộ tính năng khác có thể thu đư ợc các kết quả thử nghiệm khác. Có một số tính năng mà chúng tôi đã xem xét để dự đoán như ng vẫn chư a thể kết hợp, ví dụ nếu bộ phim đư ợc đề cử giải thư ởng của viện hàn lâm, thì nhóm sản xuất có kết nối tốt như thế nào với thế giới? Có, chúng tôi đã đư a vào các tính năng phổ biến như ng câu hỏi rộng hơ n vẫn là, ngân sách đư ợc quản lý như thế nào? Bao nhiêu đã đư ợc phân bổ cho tiếp thị và chiến dịch?. Do đó, việc đư a thêm nhiều tính năng hơ n sau khi tham khảo ý kiến của các chuyên gia trong ngành là một phần trong kế hoạch tư ơ ng lai của chúng tôi.

Độ tin cậy của công cụ và phư ơ ng pháp: Mặc dù chúng tôi đã sử dụng các công cụ và phư ơ ng pháp chuẩn phù hợp cho các nghiên cứu như vậy, như ng vẫn có khả năng tồn tại những điểm không chính xác mà chúng tôi không tính đến.

Ví dụ, chúng tôi cho rằng các thẻ bộ gen đư ợc tính toán hoàn hảo và cung cấp biểu diễn tốt hơ n về nội dung phim. Tuy nhiên, do đư ợc lấy từ thuật toán ML, các thẻ này có thể không chính xác trong một số trư ờng hợp.

Tính hợp lệ bên ngoài: Chúng tôi đồng ý rằng có một số hạn chế có khả năng ảnh hư ởng đến khả năng khái quát hóa các phát hiện của chúng tôi. Đầu tiên, vấn đề lấy mẫu phim; để giải thích rõ hơ n, chúng tôi nhận ra rằng thời thế đã thay đổi kể từ đầu những năm 1920 và do đó chúng tôi chỉ cần xem xét một khung thời gian hạn chế trư ớc thời kỳ mục tiêu, điều này dẫn đến việc giảm số điểm dữ liệu và do đó giảm số lần kiểm tra tính hợp lệ mà chúng tôi có thể thực hiện.

Hơ n nữa, chỉ có một thuật toán học máy đư ợc sử dụng, cụ thể là Random Forest. Chúng tôi đã cân nhắc sử dụng mạng nơ -ron như ng do thiếu điểm dữ liệu nên mô hình của chúng tôi sẽ bị lỗi.

6 KẾT LUẬN VÀ CÔNG VIỆC TƯ Ơ NG LAI

Trong bài báo này, chúng tôi tìm cách trả lời câu hỏi các công ty sản xuất có thể dự đoán lợi nhuận trong tư ơ ng lai của các dự án điện ảnh sắp tới như thế nào, cố gắng giúp các nhà làm phim hiểu rõ hơ n về khả năng thư ơ ng mại của các dự án của họ. Để đạt đư ợc điều này, chúng tôi đã trích xuất các tính năng trong các danh mục Nội dung, Quảng cáo, Ngày phát hành, Tài chính , Nhà sản xuất, Biên kịch, Đạo diễn, Nhà sản xuất, Diễn viên chính và Nhân viên hỗ trợ. Cuối cùng, chúng tôi đã xây dựng một mô hình phân loại rừng ngẫu nhiên, sau khi điều chỉnh siêu tham số kỹ lư ỡng, có thể dự đoán thành công lợi nhuận của các bộ phim có điểm AUC là

78%.

Công việc trong tư ơ ng lai: Ư u tiên hàng đầu là thêm nhiều tính năng hơ n và mở rộng tập dữ liệu của chúng tôi để bao gồm nhiều điểm dữ liệu hơ n cho các tính năng khả dụng. Điều này bao gồm việc thu thập dữ liệu từ các trang web truyền thông xã hội để tìm các kết nối ẩn và mạng lư ới sâu trư ớc đây chư a đư ợc biết đến, ví dụ, xây dựng biểu đồ kết nối truyền thông xã hội để xác định tốt hơ n mức độ phổ biến và các vòng tròn bên trong của nhóm sản xuất, từ đó cung cấp các dự đoán tốt hơ n cho các vấn đề cốt lõi của chúng tôi. (Các vòng tròn xã hội xác định mức độ hiện diện của một thành viên duy nhất có sức mạnh như thế nào - sự hiện diện càng tốt, lợi nhuận càng cao). Sau khi thực hiện xong điều đó, chúng tôi muốn tập trung nỗ lực vào việc sử dụng các thuật toán học tập khác để có đư ợc các mô hình mới có thể so sánh hoặc kết hợp với mô hình hiện tại của chúng tôi. Một thuật toán như vậy có thể là mạng nơ -ron vì chúng ta có thể khám phá thêm các mối quan hệ trong thế giới thực hiện tại chư a đư ợc biết đến. Tất nhiên, điều này có thể thực hiện đư ợc khi chúng tôi đã mở rộng tập dữ liệu của mình một cách đầy đủ.

TÀI LIỆU THAM KHẢO

[1] Breiman, L. Rừng ngẫu nhiên. Học máy 45, 1 (2001), 5–32.

[2] Harper, FM và Konstan, JA Bộ dữ liệu movielens: Lịch sử và bối cảnh. ACM Trans. Tư ơ ng tác. Trí tuệ. Hệ thống. 5, 4 (tháng 12/2015).

[3] Kaggle. Bộ dữ liệu phim. Trực tuyến, 2017.

[4] Kraskov, A., Stögbauer, H., và Grassberger, P. Ư ớc tính thông tin lẫn nhau. Đánh giá vật lý E 69, 6 (2004), 066138.

[5] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., và Zeileis, A. Tầm quan trọng của biến có điều kiện đối với rừng ngẫu nhiên. BMC bioinformatics 9, 1 (2008), 307.

[6] Vig, J., Sen, S., và Riedl, J. Bộ gen thẻ: Mã hóa kiến thức cộng đồng để hỗ trợ tư ơ ng tác mới. ACM Trans. Interact. Intell. Syst. 2, 3 (tháng 9 năm 2012).