Dự đoán doanh thu phim Sử dụng mô hình học máy

Vikranth Udandarao
Khoa Khoa học và Kỹ thuật Máy tính
IIIT-Delhi, Ấn Độ
vikranth22570@iiitd.ac.in

Khoa Khoa học
& Kỹ thuật Máy tính Pratyush Gupta
IIIT-Delhi, Ấn Độ
pratyush22375@iiitd.ac.in

Tóm tắt—Trong ngành công nghiệp điện ảnh đương đại, việc dự đoán chính xác doanh thu của một bộ phim là tối quan trọng để tối đa hóa lợi nhuận. Dự án này nhằm mục đích phát triển một mô hình học máy để dự đoán doanh thu của bộ phim dựa trên các tính năng đầu vào như tên phim, xếp hạng MPAA của bộ phim, thể loại phim, năm phát hành phim, Xếp hạng IMDb, số phiếu bầu của người xem, đạo diễn, biên kịch và dàn diễn viên chính, quốc gia sản xuất bộ phim, kinh phí của bộ phim, công ty sản xuất và thời lượng của bộ phim. Thông qua phương pháp có cấu trúc liên quan đến thu thập dữ liệu, xử lý trước, phân tích, lựa chọn mô hình, đánh giá và cải tiến, một mô hình dự đoán mạnh mẽ đã được xây dựng. Hồi quy tuyến tính, Cây quyết định, Hồi quy rừng ngẫu nhiên, Bagging, XGBoosting và Gradient Boosting đã được đào tạo và thử nghiệm.

Các chiến lược cải tiến mô hình bao gồm điều chỉnh siêu tham số và xác thực chéo. Mô hình kết quả cung cấp độ chính xác và khái quát đầy hứa hẹn, tạo điều kiện cho việc ra quyết định sáng suốt trong ngành công nghiệp phim ảnh để tối đa hóa lợi nhuận.

I. GIỚI THIỆU

A. Động lực

Hãy tưởng tượng bạn là một nhà làm phim hoặc giám đốc một công ty sản xuất phim và bạn có một câu hỏi lớn: điều gì làm cho một bộ phim trở thành bom tấn hay thất bại?

Bạn có thể nghĩ rằng điều đó phụ thuộc vào sức hút của các diễn viên, tầm nhìn của đạo diễn, kinh phí sản xuất hoặc thể loại của câu chuyện.

Hoặc bạn có thể nghĩ rằng chỉ có chất lượng kể chuyện mới thu hút được khán giả và đạt được tỷ lệ đánh giá cao. Nhưng câu trả lời không hề đơn giản hay dễ dàng.

Có nhiều yếu tố ảnh hưởng đến doanh thu của một bộ phim và sự kết hợp thực sự của các yếu tố này vẫn chưa được nắm vững. Đó là lý do tại sao chúng tôi đã phát triển một mô hình học máy tiết lộ các yếu tố quan trọng nhất để thành công tại phòng vé bằng cách phân tích dữ liệu thực tế từ nhiều loại phim được sản xuất trên khắp thế giới. Với mô hình của chúng tôi, các nhà làm phim có thể đưa ra quyết định sáng suốt hơn và tối ưu hóa quá trình sản xuất phim của họ để có lợi nhuận và mức độ phổ biến tối đa.

B. Cơ sở lý luận

Chúng tôi đưa ra giả thuyết rằng một số thông số có ý nghĩa hơn trong việc dự đoán doanh thu phim so với những thông số khác. Cụ thể, chúng tôi phỏng đoán rằng thành tích của đạo diễn và thể loại phim có trọng số đáng kể trong mô hình dự đoán này.

Quan sát của chúng tôi cho thấy rằng mặc dù xếp hạng IMDb thấp hơn, các bộ phim hành động thường thể hiện hiệu suất mạnh mẽ tại phòng vé. Ngược lại, các thể loại như phim hài hoặc phim chính kịch tình cảm, mặc dù có xếp hạng IMDb cao hơn, có thể không đạt được kết quả doanh thu tương đương với các phim hành động.

Những hiểu biết sâu sắc này nhấn mạnh sự tương tác phức tạp giữa các thuộc tính của phim và sở thích của khán giả, thúc đẩy chúng ta coi trọng hơn các yếu tố như lịch sử đạo diễn và phân loại thể loại trong khuôn khổ dự đoán của mình.

C. Tổng quan

Trong dự án này, chúng tôi áp dụng phương pháp có cấu trúc để xây dựng và đánh giá mô hình dự đoán của mình. Đầu tiên, chúng tôi thu thập một tập dữ liệu lớn về phim và các tính năng của chúng từ nhiều nguồn khác nhau và tùy chỉnh các tập dữ liệu cho phù hợp với nhu cầu của chúng tôi.

Sau đó, chúng tôi xử lý trước dữ liệu để xử lý các giá trị bị thiếu, giá trị ngoại lai và các biến phân loại. Chúng tôi thực hiện phân tích dữ liệu để khám phá dữ liệu và hiểu các đặc điểm và mối quan hệ của nó. Chúng tôi sử dụng thống kê mô tả, thống kê suy luận và các kỹ thuật trực quan hóa dữ liệu để có được thông tin chi tiết về dữ liệu như sử dụng biểu đồ để so sánh độ chính xác của hiệu suất dữ liệu đào tạo và thử nghiệm của mô hình.

Sau đó, chúng tôi chọn một số thuật toán học máy phù hợp với các tác vụ hồi quy, chẳng hạn như cây quyết định và rừng ngẫu nhiên. Chúng tôi đào tạo và kiểm tra các mô hình của mình bằng cách sử dụng xác thực chéo và so sánh hiệu suất của chúng bằng các số liệu như lỗi trung bình R bình phương và Lỗi phần trăm tuyệt đối trung bình. Chúng tôi cũng áp dụng các chiến lược cải tiến mô hình như điều chỉnh siêu tham số, lựa chọn tính năng và chính quy hóa để nâng cao độ chính xác và khái quát hóa của các mô hình của chúng tôi. Mô hình kết quả cung cấp các kết quả đầy hứa hẹn và có thể được sử dụng để dự đoán doanh thu của bất kỳ bộ phim nào dựa trên các tính năng của nó.

II. TỔNG QUAN TÀI LIỆU

Trong phần này, chúng ta sẽ xem định nghĩa của Principal Com-ponent Analysis (PCA), Label Encoder, SelectKBest features, GridSearchCV, Train Test Split và cung cấp một số nội dung tài liệu về các mô hình mà chúng ta sẽ sử dụng. Chúng tôi cũng sẽ giải thích các số liệu đánh giá 'Điểm R^2' và 'MAPE'.

Phân tích thành phần chính (PCA)

PCA là một thủ tục thống kê sử dụng phép biến đổi trực giao để chuyển đổi một tập hợp các quan sát của các biến có thể tương quan thành một tập hợp các giá trị không tương quan tuyến tính.

biến được gọi là thành phần chính. Kỹ thuật này được sử dụng để nhấn mạnh sự thay đổi và đưa ra các mô hình mạnh mẽ trong một tập dữ liệu.

Bộ mã hóa nhãn

Label Encoder là một phương pháp tiện ích để chuyển đổi dữ liệu phân loại thành dữ liệu số. Nó gán mỗi danh mục duy nhất trong dữ liệu cho một giá trị số nguyên, làm cho dữ liệu phù hợp hơn cho xử lý thuật toán.

Tính năng SelectKBest

SelectKBest là phương pháp chọn tính năng trong Scikit-Learn. Phương pháp này chọn các tính năng theo k điểm cao nhất của một hàm tính điểm cụ thể. Đây là cách để chọn 'k' tính năng tốt nhất trong tập dữ liệu của bạn, trong đó 'k' là tham số bạn chọn.

LướiTìmKiếmCV

GridSearchCV là một hàm thư viện là thành viên của gói lựa chọn mô hình sklearn. Nó giúp lặp qua các siêu tham số được xác định trước và phù hợp với bộ ước tính (mô hình) của bạn trên tập huấn luyện của bạn. Ngoài ra, bạn có thể chỉ định số lần xác thực chéo cho mỗi tập siêu tham số.

Train Test Split

Các chức năng từ 'train test split' module 'sklearn.model selection' là một tiện ích chia một tập dữ liệu thành các tập con huấn luyện và thử nghiệm ngẫu nhiên. Mỗi tập con là riêng biệt, nghĩa là không có điểm dữ liệu nào có thể có trong cả hai tập con. Điều này cho phép mô hình được huấn luyện trên một tập con của dữ liệu, sau đó được xác thực trên một tập con hoàn toàn riêng biệt. Trong trường hợp của chúng tôi, chúng tôi đã áp dụng phân chia 80/20 trên tập dữ liệu của mình để huấn luyện và thử nghiệm các mô hình của chúng tôi.

Mô hình

A. Hồi quy tuyến tính

Hồi quy tuyến tính là một phương pháp thống kê để mô hình hóa mối quan hệ giữa biến phụ thuộc và một hoặc nhiều biến độc lập.

B. Cây quyết định

Cây quyết định là một công cụ hỗ trợ quyết định sử dụng mô hình dạng cây của các quyết định và hậu quả có thể xảy ra của chúng. Đây là một cách để hiển thị thuật toán chỉ chứa các câu lệnh điều khiển có điều kiện.

C. Tăng cường độ dốc

Gradient Boosting là một kỹ thuật học máy dành cho các bài toán hồi quy và phân loại, tạo ra mô hình dự đoán dưới dạng tập hợp các mô hình dự đoán yếu, thường là cây quyết định.

D. Bagging

Bootstrap Aggregating, thường được viết tắt là Bagging, là một siêu thuật toán được thiết kế để cải thiện tính ổn định và độ chính xác của các thuật toán học máy được sử dụng trong phân loại thống kê và hồi quy. Nó cũng làm giảm phương sai và giúp tránh tình trạng quá khớp.

E. Rừng ngẫu nhiên

Rừng ngẫu nhiên là một phương pháp học tập hoạt động bằng cách xây dựng nhiều cây quyết định tại thời điểm đào tạo và đưa ra lớp là chế độ của các lớp (phân loại) hoặc dự đoán trung bình (hồi quy) của từng cây.

F. Tăng cường XG

XGBoost là một thư viện tăng cường độ dốc phân tán được tối ưu hóa, được thiết kế để có hiệu suất cao, linh hoạt và di động.
Nó triển khai các thuật toán học máy theo khuôn khổ Gradi-ent Boosting. Thư viện XGBoost cung cấp một số lợi thế so với các thuật toán học máy truyền thống, bao gồm: • Song song hóa: XGBoost hỗ trợ song song hóa, cho phép nó đào tạo các

mô hình hiệu quả trên các tập dữ liệu lớn bằng cách tận dụng nhiều lõi CPU hoặc GPU. • Chính quy hóa: XGBoost bao gồm các kỹ thuật chính quy hóa tích hợp, chẳng hạn như chính quy hóa L1 và L2, để

ngăn chặn tình trạng quá khớp và cải thiện tổng quát hóa mô hình. • Xử lý dữ liệu bị thiếu: XGBoost có thể tự động xử lý dữ liệu bị thiếu mà không cần phải suy diễn, giúp nó mạnh mẽ và hiệu quả hơn.

Số liệu đánh giá

Điểm R^2

Điểm R^2, còn được gọi là hệ số xác định, là một phép đo thống kê cho thấy tỷ lệ phương sai của một biến phụ thuộc được giải thích bởi một biến độc lập hoặc các biến trong mô hình hồi quy. Nó cung cấp một dấu hiệu về mức độ phù hợp và do đó là một phép đo về mức độ các mẫu chưa thấy có khả năng được mô hình dự đoán tốt như thế nào.

Sai số phần trăm tuyệt đối trung bình (MAPE)

Sai số phần trăm tuyệt đối trung bình (MAPE) là một phép đo thống kê được sử dụng để đánh giá độ chính xác của phương pháp dự báo trong các nghiên cứu dự đoán. Đây là giá trị trung bình của tất cả các sai số phần trăm tuyệt đối giữa giá trị dự đoán và giá trị thực tế. Nó cung cấp sự hiểu biết về sai số dự đoán theo tỷ lệ phần trăm của các giá trị thực tế. Giá trị MAPE thấp hơn cho thấy dữ liệu phù hợp hơn. MAPE cũng có thể được hiểu là nghịch đảo của độ chính xác của mô hình, nhưng cụ thể hơn là chênh lệch phần trăm trung bình giữa các dự đoán và mục tiêu dự kiến của chúng trong tập dữ liệu. Ví dụ, nếu MAPE của bạn là 10% thì dự đoán của bạn trung bình cách giá trị thực tế mà chúng hướng tới 10%.

III. LỰA CHỌN TẬP DỮ LIỆU

A. Sửa đổi Chiến lược Bộ dữ liệu Trong đề

xuất ban đầu của chúng tôi, chúng tôi đã xây dựng một bộ dữ liệu riêng bằng cách tích hợp bốn bộ dữ liệu riêng biệt: Bộ dữ liệu Ngành công nghiệp Phim ảnh , Bộ dữ liệu phim nhiều thể loại IMDb 5000, Bộ dữ liệu phim IMDb 5000, và Top 500 Phim có kinh phí cao nhất.
Sự tích hợp này được thực hiện tuần tự, tạo ra một tập dữ liệu toàn diện bao gồm 7118 phim với

các biến đầu vào như ngân sách, đạo diễn, thể loại, đoàn làm phim chính, và xếp hạng IMDb, và biến đầu ra là tổng doanh thu. (Tham khảo README để có lời giải thích chi tiết về quá trình xây dựng tập dữ liệu).

Tuy nhiên, trong quá trình thực hiện dự án, chúng tôi nhận thấy cần phải thay đổi so với bộ dữ liệu ban đầu được đề xuất. Chúng tôi đã chọn một tập dữ liệu tinh chỉnh có nguồn gốc hoàn toàn từ The Movies Bộ dữ liệu ngành, đạt được bằng cách loại trừ các mục nhập bị thiếu giá trị. Quyết định quan trọng này được thông báo bởi vô số những cân nhắc được trình bày chi tiết trong các tiểu mục sau.

B. Cơ sở lý luận đằng sau quá trình chuyển đổi tập dữ liệu

Thách thức chính gặp phải với tập dữ liệu hoàn thiện ban đầu của chúng tôi là hiệu suất không tối ưu của nó với các mô hình dự đoán, phản ánh ở tỷ lệ chính xác thấp hơn. Điều này là được quy cho cách tiếp cận ban đầu của chúng tôi trong việc lựa chọn các biến đầu vào không có sự phân tích kỹ lưỡng các dữ liệu có sẵn, dẫn đến chúng tôi ép buộc các tập dữ liệu vào một khuôn khổ được hình thành trước của các yếu tố quyết định doanh thu, do đó làm tổn hại đến tính toàn vẹn của tập dữ liệu cuối cùng. Ngoài ra, khả năng gây nhầm lẫn do các bộ phim có tiêu đề tương tự nhau đòi hỏi phải thận trọng cách tiếp cận để hợp nhất tập dữ liệu, trong đó trình bày tập hợp riêng của nó sự phức tạp.

Nhận thức được những vấn đề này, chúng tôi đã dừng lại để đánh giá lại phương pháp luận. Chúng tôi đã đi sâu vào việc kiểm tra sâu hơn dữ liệu xung quanh dự đoán doanh thu phim và thu thập thông tin chi tiết vào việc cấu trúc tập dữ liệu để đào tạo một mô hình vừa toàn diện và mạnh mẽ trong việc dự báo thu nhập từ phim ảnh.

C. Lợi ích của Bộ dữ liệu được tối ưu hóa

Việc chúng tôi tiếp tục khám phá một tập dữ liệu lý tưởng bao gồm cốt lõi của nghiên cứu đã dẫn chúng tôi đến "Tập dữ liệu phim ảnh", phù hợp hoàn hảo với tiêu chí của chúng tôi. Chúng tôi đảm bảo độ tin cậy của tập dữ liệu bằng cách loại bỏ các mục có giá trị null trước khi đưa nó vào các mô hình dự đoán của chúng tôi. Một điều quan trọng lợi thế của tập dữ liệu hiện tại của chúng tôi là khả năng giải quyết những hạn chế của dữ liệu trước đây của chúng tôi. Nó bao gồm các biến đầu vào bổ sung như Năm, Công ty sản xuất và Phiếu bầu (cùng với điểm số IMDb), cùng nhau nâng cao độ chính xác của mô hình học máy của chúng tôi trong việc dự đoán phim doanh thu. Hơn nữa, việc lấy nguồn dữ liệu từ một nguồn duy nhất đã loại bỏ tiếng ồn bên ngoài phát sinh trước đó hợp nhất tập dữ liệu.

IV. PHÂN TÍ CH DỮ LIỆU

Trong giai đoạn phân tích dữ liệu, chúng tôi sẽ kiểm tra kỹ lưỡng bộ dữ liệu phim được thu thập để hiểu rõ hơn về cấu trúc của nó, phân phối và mối quan hệ giữa các tính năng và mục tiêu biến (thu nhập từ phim). Chúng tôi sẽ sử dụng thống kê mô tả, thống kê suy luận và kỹ thuật trực quan hóa dữ liệu để khám phá dữ liệu.

A. Thống kê mô tả

Chúng tôi đã tính toán các số liệu thống kê tóm tắt như trung bình, trung vị, độ lệch chuẩn, tối thiểu, tối đa và tứ phân vị cho các tính năng số như ngân sách và xếp hạng. Điều này đã cho chúng tôi một hiểu biết chung về xu hướng trung tâm và sự lan truyền của dữ liệu.

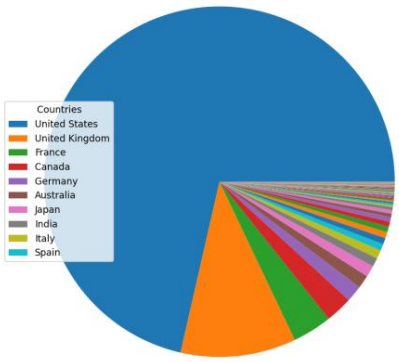
BẢNG I: Mô tả các tính năng của bộ dữ liệu phim

Tính năng Tên	Sự miêu tả
Tên	Tên của bộ phim
Xếp hạng	Xếp hạng MPAA của bộ phim
Thể loại	Thể loại phim
Năm	Năm bộ phim được phát hành
Phát hành	Ngày phát hành của bộ phim
Điểm	Xếp hạng IMDb của bộ phim
Phiếu bầu	Số lượng phiếu bầu mà bộ phim nhận được
Giám đốc	Người đạo diễn bộ phim
Nhà văn	Người viết kịch bản phim
Ngôi sao	Nam diễn viên chính trong phim
Quốc gia	Quốc gia nơi bộ phim được sản xuất
Ngân sách	Ngân sách của bộ phim
Công ty	Công ty sản xuất bộ phim
Thời gian chạy	Thời lượng của bộ phim

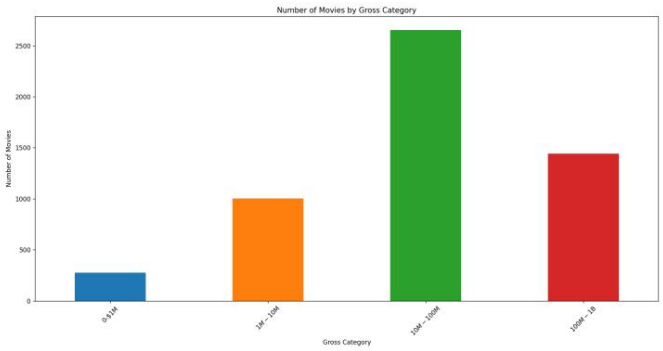
B. Thống kê suy luận

Chúng tôi sẽ tiến hành các thử nghiệm thống kê để phân tích mối quan hệ giữa các tính năng khác nhau và biến mục tiêu. Ví dụ, chúng ta có thể sử dụng phân tích tương quan để kiểm tra sức mạnh và hướng của các mối quan hệ tuyến tính giữa các tính năng số và thu nhập.

V. TRỰC QUAN HÓA DỮ LIỆU



Hình 1: Phân phối phim theo quốc gia



Hình 2: Biểu đồ Histogram của các danh mục tổng

VI.XỬ LÝ TRƯỚC

A. Các tham số của tập dữ liệu

Chúng ta có 14 tham số, bao gồm cả kiểu dữ liệu số và không phải số. 1) tên 2) xếp hạng 3) thể loại 4) năm 5) phát hành 6) điểm 7) phiếu bầu 8) đạo diễn 9) biên kịch 10) ngôi sao 11) quốc gia 12) ngân sách 13) công ty 14) thời gian chạy

Chúng ta có 9 kiểu dữ liệu không phải số 1) tên 2) xếp hạng 3) thể loại 4) phát hành 5) đạo diễn 6) biên kịch 7) ngôi sao 8) quốc gia 9) công ty

Để chuẩn bị dữ liệu cho các mô hình hồi quy, chúng ta cần mã hóa các tính năng không phải số thành nhãn số. Chúng ta sẽ sử dụng LabelEncoder từ scikit-learn cho mục đích này.

NhãnEncoder

LabelEncoder là một công cụ tiện dụng để chuẩn hóa và chuyển đổi các nhãn không phải dạng số thành dạng số tương đương.
Điều quan trọng cần lưu ý là các nhãn phải có thể băm và so sánh được để quá trình này hoạt động hiệu quả. Chúng ta sẽ bắt đầu bằng cách lắp bộ mã hóa nhãn và lấy các nhãn đã mã hóa để xử lý thêm.

name	0
rating	77
genre	0
year	0
released	2
score	3
votes	3
director	0
writer	3
star	1
country	3
budget	2171
gross	189
company	17
runtime	4
dtype: int64	

Hình 3: Giá trị Null

Xử lý giá trị Null

Trong Bộ dữ liệu ngành công nghiệp phim ảnh, có 2.247 giá trị null trong 11 tham số, tổng cộng là 7669. Vì ngân sách và tổng doanh thu là tham số chính và đầu ra của chúng tôi nên chúng tôi đã loại bỏ các tập dữ liệu đó và còn lại 5422 tập dữ liệu.

PCA

Chúng tôi đã thử sử dụng Phân tích thành phần chính nhưng không thể đạt được mức tăng đáng kể về độ chính xác. Chúng tôi tin rằng lý do cho điều này là PCA hoạt động hiệu quả khi có nhiều tham số và tập dữ liệu hơn. Nó sẽ hữu ích hơn trong hồi quy/phân loại hình ảnh hoặc âm thanh vì các kích thước của hình ảnh/âm thanh sẽ được đưa vào sử dụng. Do đó, chúng tôi đã quyết định loại bỏ PCA trong quá trình tiền xử lý của mình.

Tính năng tốt nhất

Chúng tôi muốn biết những tính năng nào đóng góp nhiều hơn cho doanh thu của bộ phim.

Vì vậy, chúng tôi đã sử dụng SelectKBest , sau đó phân loại và dự đoán tất cả 8668 tính năng có thể xem trong Điểm tính năng.

Vì có rất nhiều tính năng để chúng tôi phân tích nên chúng tôi chỉ in những tính năng có điểm lớn hơn 100 để xem trong Điểm tính năng tốt nhất.

Feature	Score
rating_PG-13	194.402250
rating_R	337.959444
genre_Action	239.210623
genre_Animation	276.657909
genre_Comedy	116.786195
director_Anthony Russo	238.515051
director_James Cameron	127.387274
writer_Christopher Markus	199.157551
writer_James Cameron	108.337247
star_Chris Pratt	105.223686
star_Daisy Ridley	120.605232
star_Daniel Radcliffe	102.613119
star_Robert Downey Jr.	151.485500
company_Lucasfilm	110.297606
company_Marvel Studios	496.764152
company_Pixar Animation Studios	107.263914
company_Walt Disney Pictures	172.259012
year	440.975532
score	282.397728
votes	3292.085413
budget	6569.008340
runtime	446.121279

Hình 4: K Tính năng tốt nhất

Theo trực giác của chúng tôi, ngân sách đóng vai trò quyết định nhất, với số điểm khoảng 6569. Tuy nhiên, điều bất ngờ là phiếu bầu lại đóng vai trò quan trọng thứ hai trong việc dự đoán doanh thu.

Ngoài ra, thời gian chạy, năm và điểm số cũng đóng vai trò khá tốt vai trò trong việc dự đoán doanh thu.

Ngoài ra, còn có một số công ty và cá nhân trong lĩnh vực đạo diễn, diễn viên và sản xuất đã tác động đến doanh thu phòng vé của bộ phim.

Ngoài ra, có một xu hướng chung có thể nhận thấy là phim được xếp loại PG-13 và R có xu hướng bán chạy.

Hành động, Hoạt hình và Hài kịch là những thể loại có thành tích tốt và điều này cũng có xu hướng làm tăng doanh thu của bộ phim.

VII. LỰA CHỌN MÔ HÌNH

Để dự đoán doanh thu phim, chúng tôi đã khám phá một số kỹ thuật hồi quy và phương pháp tổng hợp như Linear Re-gression có và không có PCA Preprocessing, Decision Trees, Gradient Boosting, Bagging, Random Forests và XGBoost-ing. Các phương pháp tổng hợp như Random Forests và Gradient Boosting có khả năng xử lý các tương tác phức tạp và phi tuyến tính trong dữ liệu.

1) Hồi quy tuyến tính: Ban đầu được sử dụng để khám phá các mối quan hệ tuyến tính của tập dữ liệu và xác định xem mô hình đường thẳng có thể biểu diễn dữ liệu một cách đầy đủ hay không.

2) Gradient Boosting: Được sử dụng như một kỹ thuật dự đoán để tạo ra một mô hình tổng hợp bao gồm các mô hình dự đoán yếu hơn (thường là cây quyết định). Mỗi người học tiếp theo trong tổng hợp đều hướng đến việc sửa lỗi do người học trước đó mắc phải.
GB được biết đến với khả năng chống lại hiện tượng quá khớp, khiến nó trở thành một thuật toán có giá trị cho các tác vụ dự đoán.

3) Rừng ngẫu nhiên: Được sử dụng như một phương pháp học tập tổng hợp để phân loại. RF xây dựng nhiều cây quyết định và đưa ra dự đoán dựa trên số phiếu bầu đa số của các cây này. Mặc dù đã nỗ lực giảm chiều dữ liệu, mô hình RF vẫn thể hiện sự quá khớp trên tập huấn luyện và không vượt trội hơn các mô hình khác đáng kể.

4) Cây quyết định: Được sử dụng như một mô hình hỗ trợ quyết định phân cấp. DT sử dụng cấu trúc giống như cây để biểu diễn các quyết định và hậu quả của chúng, bao gồm các sự kiện ngẫu nhiên, chi phí tài nguyên và tiện ích. Chúng cung cấp hình ảnh trực quan rõ ràng về quá trình ra quyết định của thuật toán.

5) Bagging: Được sử dụng để cải thiện hiệu suất mô hình bằng cách đào tạo nhiều mô hình trên các tập hợp con ngẫu nhiên của dữ liệu gốc.
Kết quả từ các mô hình này được kết hợp thông qua cơ chế bỏ phiếu để đưa ra dự đoán, mang lại những dự đoán chính xác và đáng tin cậy hơn.

6) XGBoost: Được áp dụng như một thuật toán học máy phổ biến trong học tập tổng hợp. XGBoost hiệu quả cho các tác vụ học có giám sát như hồi quy và phân loại. Nó xây dựng mô hình dự đoán theo từng bước bằng cách kết hợp các dự đoán từ nhiều mô hình riêng lẻ, thường là cây quyết định, để tăng cường độ chính xác của dự đoán và hiệu suất mô hình.

VIII. CẢI TIẾN MÔ HÌNH

Trong nỗ lực nâng cao hiệu suất của mô hình, chúng tôi sẽ sử dụng một số chiến lược. Hãy cùng phân tích các thành phần chính:

A. Chuẩn hóa dữ liệu với Standard Scaler

Để đảm bảo tính nhất quán giữa các tính năng, chúng tôi sẽ sử dụng Standard Scaler. Kỹ thuật này chuẩn hóa từng tính năng bằng cách trừ giá trị trung bình và chia cho độ lệch chuẩn.
Bằng cách áp dụng phép biến đổi này, chúng ta đưa tất cả các tính năng về cùng một tỷ lệ, điều này có thể cải thiện đáng kể hiệu suất của mô hình.

B. Kỹ thuật tính năng: Biến đổi logarit

Chúng ta sẽ tập trung vào hai cột cụ thể: 'budget' và 'gross' doanh thu. Các cột này thường biểu hiện phân phối lệch. Để giải quyết vấn đề này, chúng ta sẽ áp dụng phép biến đổi logarit. Chuyển đổi 'budget' và 'gross' thành log(budget) và log(gross) bằng cách sử dụng numpy.log1p, tương ứng, sẽ ổn định phương sai của chúng và tăng cường độ mạnh mẽ của mô hình. Chúng tôi cũng sẽ báo cáo MSLE khi phần trên được triển khai để kiểm tra xem độ lệch của mô hình có thấp không.

C. Điều chỉnh siêu tham số bằng GridSearchCV

Bây giờ, chúng ta hãy đi sâu vào chi tiết về điều chỉnh siêu tham số. GridSearchCV là một công cụ mạnh mẽ thực hiện tìm kiếm toàn diện trên các giá trị tham số được chỉ định cho một bộ ước tính. Trong trường hợp của chúng tôi, chúng tôi đang sử dụng Gradient Boosting Regressor với hàm mất lỗi bình phương. Sau đây là cách thức hoạt động:

- Chúng tôi định nghĩa một từ điển có tên là param_grid chứa nhiều siêu tham số khác nhau (ví dụ: số lượng ước lượng, độ sâu tối đa và tốc độ học). • GridSearchCV kiểm tra một cách có hệ thống các tổ hợp khác nhau của các siêu tham số này bằng cách sử dụng xác thực chéo (trong trường hợp của chúng tôi là xác thực chéo 5 lần).

- Sự kết hợp tốt nhất của các tham số, được xác định bởi điểm R bình phương cao nhất, được trích xuất bằng cách sử dụng thuộc tính best_params. • Được trang bị các tham số tối ưu này, chúng tôi khởi tạo mô hình cuối cùng.

- Sau đó, mô hình được đào tạo trên dữ liệu đào tạo và được đánh giá trên cả tập dữ liệu đào tạo và tập dữ liệu thử nghiệm để xác thực độ chính xác của nó.

Bằng cách thực hiện quy trình này, chúng tôi mong muốn tạo ra một mô hình mạnh mẽ và hiệu suất cao.

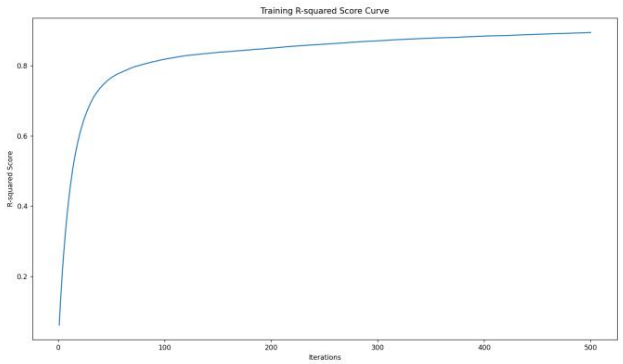
D. Triển khai XGBRegressor bằng thư viện xgboost

XGBoost (Extreme Gradient Boosting) là một triển khai mạnh mẽ và hiệu quả của thuật toán tăng cường gradient, được sử dụng rộng rãi cho nhiều tác vụ học máy khác nhau, bao gồm hồi quy và phân loại. Trong dự án này, chúng tôi đã sử dụng XGBRegressor từ thư viện xgboost để đào tạo và kiểm tra mô hình hồi quy nhằm dự đoán doanh thu phim dựa trên nhiều tính năng khác nhau như ngày phát hành, thể loại, đạo diễn, v.v.

Để đào tạo và đánh giá mô hình hồi quy XGBoost, chúng tôi đã làm theo các bước sau:

1) Tải tập dữ liệu phim và xử lý trước dữ liệu bằng mã hóa các tính năng phân loại bằng cách sử dụng LabelEncoder từ học scikit.
2) Chia tập dữ liệu thành các tập huấn luyện và thử nghiệm bằng cách sử dụng train_test_split từ scikit-learn.
3) Xác định một lớp gọi lại tùy chỉnh TrackR2Score kế thừa từ xgb.callback.TrainingCallback. Lớp này ghi đè phương pháp lặp lại sau để tính toán và lưu trữ điểm R-squared trên tập huấn luyện sau mỗi lần sự lặp lại trong quá trình đào tạo.
4) Thực hiện điều chỉnh siêu tham số bằng GridSearchCV từ scikit-learn. Chúng tôi đã vượt qua lệnh gọi lại TrackR2Score đến trường hợp XGBRegressor được sử dụng làm trình ước tính trong GridSearchCV, cho phép chúng tôi theo dõi điểm R-squared trong quá trình tìm kiếm lưới.
5) Đào tạo mô hình XGBRegressor cuối cùng bằng cách sử dụng tốt nhất siêu tham số được tìm thấy bởi GridSearchCV.
6) Đánh giá hiệu suất của mô hình trên cả hai phương pháp đào tạo và thử nghiệm các bộ dữ liệu sử dụng nhiều số liệu khác nhau, chẳng hạn như R-bình phương điểm và Sai số phần trăm tuyệt đối trung bình (MAPE).
7) Hình dung giá trị thực tế so với giá trị dự đoán cho cả hai bộ đào tạo và thử nghiệm.
8) Vẽ đường cong điểm R bình phương đào tạo để quan sát cải tiến dần dần mô hình trong quá trình đào tạo quá trình.

Biểu đồ sau đây cho thấy đường cong điểm R^2 đào tạo, minh họa sự cải thiện dần dần của mô hình trong quá trình đào tạo quá trình: Bằng cách tận dụng thư viện XGBoost mạnh mẽ và



Hình 5: Đường cong điểm R^2 đào tạo

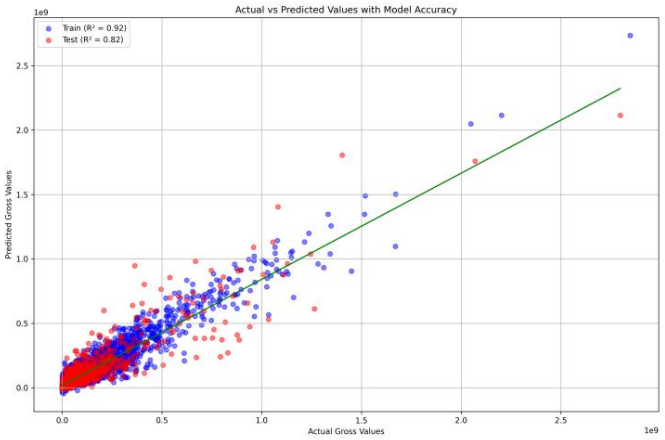
triển khai một hàm gọi lại tùy chỉnh, chúng tôi đã có thể đào tạo và đánh giá hiệu quả một mô hình hồi quy để dự đoán doanh thu phim trong khi theo dõi hiệu suất của mô hình trong quá trình đào tạo.

IX. ĐÁNH GIÁ MÔ HÌNH

Chúng tôi sẽ đánh giá hiệu suất của các mô hình của chúng tôi bằng cách sử dụng các số liệu thích hợp như Sai số phần trăm tuyệt đối trung bình (MAPE) và Hệ số xác định (R^2). Khi kết hợp với nhau, hai số liệu này cung cấp một cái nhìn toàn diện về hiệu suất mô hình của bạn. Điểm R^2 cho bạn biết mức độ tốt mô hình của bạn đang nắm bắt các mẫu trong dữ liệu, trong khi MAPE cung cấp cho bạn cảm giác về lỗi phần trăm trung bình trong dự đoán của mô hình của bạn. Ngoài ra, chúng tôi sẽ so sánh hiệu suất của mô hình của chúng tôi với các mô hình cơ sở để đánh giá hiệu quả.

Người mẫu	R^2	BẢN ĐỒ
Hồi quy tuyến tính		
Bộ đào tạo	0,6553	35,23%
Bộ kiểm tra	0,6706	18,49%
Cây quyết định		
Bộ đào tạo	0,8664	13,00%
Bộ kiểm tra	0,6947	4,60%
Đóng bao		
Bộ đào tạo	0,8583	13,32%
Bộ kiểm tra	0,7719	5,67%
Tăng cường độ dốc		
Bộ đào tạo	0,9158	10,57%
Bộ kiểm	0,8242	5,69%
tra eXtreme Gradient Boosting		
Bộ đào tạo	0,9079	9,70%
Bộ kiểm tra	0,8102	5,53%
Rừng ngẫu nhiên		
Bộ đào tạo	0,8728	14,29%
Bộ kiểm tra	0,7786	5,33%

BẢNG II: Kết quả đánh giá mô hình



Hình 6: Kết quả thử nghiệm tăng cường độ dốc

X. GIAO DIỆN DÒNG LỆNH

Chúng tôi đã tạo ra một giao diện dòng lệnh cho các nhà sản xuất truy cập mô hình của chúng tôi, tại đó họ sẽ được nhắc nhập 14 các thông số mà mô hình sẽ được đào tạo. Chúng tôi có tùy chọn cho nhà sản xuất để lựa chọn mô hình mà anh ta muốn để sử dụng.

1) Hồi quy tuyến tính
2) Cây quyết định
3) Đóng gói
4) Rừng ngẫu nhiên
5) Tăng tốc XG
6) Tăng cường độ dốc

CLI có thể được truy cập trong main.py Nhà sản xuất phải chạy python main.py và sau đó nhập 14 tham số để dự đoán doanh thu cho bộ phim sắp ra mắt của mình như chúng ta đã hình dung trong Bài toán.

XI. KẾT LUẬN

Tính đến thời điểm hiện tại, chúng tôi nhận thấy Gradient Boosting là mô hình tốt nhất của mình, đạt được:

• Độ chính xác đào tạo cuối cùng: 91,58% • Độ chính xác thử nghiệm cuối cùng: 82,42% Có thể truy cập ứng dụng trong main.py Có thể xem mô hình trong Models/gradient boost.py.
Các mô hình khác có thể được xem tại đây.
Độ chính xác của tất cả các mô hình có thể được xem tại đây.

LỜI CẢM ƠN

Các tác giả xin gửi lời cảm ơn chân thành nhất đến Tiến sĩ AV Subramanyam (Khoa Khoa học và Kỹ thuật Máy tính, IIIT-Delhi) vì sự hướng dẫn vô giá của họ trong suốt dự án. Phản hồi sâu sắc và chuyên môn của họ đã đóng vai trò quan trọng trong việc định hình dự án này thành hình thức cuối cùng.

TÀI LIỆU THAM KHẢO

[1] Vr, Nithin & Pranav, M & Babu, PB & Lijiya, A.. (2014). Dự đoán thành công của phim dựa trên dữ liệu IMDB. Tạp chí quốc tế về trí tuệ kinh doanh. 003. 34-36. 10.20894/IJBI.105.003.002.004.

[2] Pradeep, Kavya & TintuRosmin, C & Durom, Sherly & Anisha, G. (2020). Thuật toán cây quyết định để dự đoán chính xác xếp hạng phim. 853-858. 10.1109/ICCMC48092.2020.ICCMC-000158.

[3] Garima Verma và Hemraj Verma, "Dự đoán thành công của phim Bollywood bằng kỹ thuật học máy", IEEE, 2019.

[4] Rijul Dhir và Anand Raj, "Dự đoán thành công của phim bằng cách sử dụng máy học và so sánh của chúng," Hội nghị quốc tế IEEE về máy tính và truyền thông mạng an toàn, 2018.

[5] Ashutosh Kanitar, "Dự đoán thành công của phim Bollywood bằng thuật toán học máy", Hội nghị quốc tế lần thứ ba của IEEE về mạch, điều khiển, truyền thông và máy tính, 2018.

[6] Wales, Lorene. (2017). Hướng dẫn đầy đủ về sản xuất phim và kỹ thuật số: Con người và quy trình. 10.4324/9781315294896.

[7] Beyza C¸ izmeci và Sule Gund ¨ uz¨ O¨ g¨ud¨ uc¨ u, "Dự đoán xếp hạng IMDb của phim trước khi phát hành bằng máy phân tích nhân tử sử dụng phương tiện truyền thông xã hội," Hội nghị quốc tế lần thứ 3 của IEEE về khoa học máy tính và kỹ thuật, 2018.

[8] Steve Shim và Mohammad Pourhomayoun, "Dự đoán doanh thu thị trường phim bằng cách sử dụng dữ liệu truyền thông xã hội," Hội nghị quốc tế IEEE về tái sử dụng và tích hợp thông tin, 2017.

[9] Nahid Quader và Md. Osman Gani, "Một phương pháp tiếp cận học máy để dự đoán doanh thu phòng vé phim," Công nghệ thông tin (ICCIT), tháng 12 năm 2017.

[10] Beyza C¸ izmeci và Sule Gund ¨ uz¨ O¨ g¨ud¨ uc¨ u, "Dự đoán xếp hạng IMDb của phim trước khi phát hành bằng máy phân tích nhân tử sử dụng phương tiện truyền thông xã hội," Hội nghị quốc tế về máy tính và công nghệ thông tin (IC-CIT), 2017.

[11] Subramaniyaswamy V., và Vignesh Vaibhav M., "Dự đoán thành công phòng vé phim bằng cách sử dụng hồi quy bội và SVM," Hội nghị quốc tế về máy tính và công nghệ thông tin (ICCIT), 2017.

[12] MH Latif, H. Afzal, "Dự đoán mức độ phổ biến của phim bằng kỹ thuật học máy", Đại học Khoa học và Công nghệ Quốc gia, H-12, ISB, tập 16, số 8, trang 127–131, 2016.

[13] DA, Olubukola & OM, Stephen & AK, Funmilayo & Omotunde, Ayokunle & A., Oyebola & Oduroye, Ayorinde & Ajayi, Wumi & Yaw, Mensah. (2021). Dự đoán thành công của phim bằng cách sử dụng khai thác dữ liệu. Tạp chí máy tính, mạng và công nghệ thông tin của Anh. 4. 22-30. 10.52589/BJCNIT-CQOCIREC.