TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN CUỐI KỲ

HỌC PHẦN: KHOA HỌC DỮ LIỆU

TÊN ĐỀ TÀI

Predicting Movie Ratings

Nhóm	5
Họ Và Tên Sinh Viên	Lớp Học Phần
Phan Khánh Ngân	20.10
Lê Văn Thịnh	
Nguyễn Phan Nhật Hoàng	

ĐÀ NẴNG, 06/2023

TÓM TẮT

Tiểu luận này tập trung vào việc giải quyết bài toán dự đoán ratings, dựa vào "pre-released features" của các bộ phim trước khi phát hành. Thông qua sử dụng các phương pháp, kỹ thuật khai phá dữ liệu và học máy bao gồm thu thập, tiền xử lý dữ liệu, xây dựng và tinh chỉnh mô hình dự đoán, có thể xây dựng hệ thống dự đoán có độ chính xác cao. Có nhiều mô hình hồi quy có thể ứng dụng cho bào toán này, trong đó có Random Forests và Support Vector Machines. Với kết quả thu được, Random Forest (regression) dự đoán cho độ chính xác tốt hơn Support Vector Machine (regression) với MAE, RMSE thấp hơn và R2 cao hơn. Kết quả này khẳng định có thể sử dụng mô hình RF vào hệ thống dự đoán ratings thực tế, giúp người sử dụng có thể đưa ra quyết định có hay không nên sản xuất một bộ phim.

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Trạng thái
Nguyễn Phan Nhật Hoàng	- Thu thập dữ liệu - Mô tả dữ liệu	Đã hoàn thành
Lê Văn Thịnh	- Xử lý dữ liệu: • Xử lý dữ liệu trống. • Biến đổi phân loại dữ liệu. • Thêm các đặc trưng mới. • Xử lý ngoại lệ. • Chuẩn hóa dữ liệu. - Trực quan hóa dữ liệu sau khi xử lý dữ liệu.	Đã hoàn thành
Phan Khánh Ngân	- Thực hiện lựa chọn đặc trưng bằng thuật toán RFE - Xây dựng hai mô hình hồi quy (RF và SVM) - Triển khai thuật toán lựa chọn tham số tối ưu - Đánh giá, so sánh hai mô hình hồi quy (RF và SVM)	Đã hoàn thành

MỤC LỤC

1. Giới thiệu

Vấn đề cần giải quyết

Trong thời gian gần đây, ngành công nghiệp phim phát triển rất mạnh mẽ trên các quốc gia trên toàn thế giới. Điều này dẫn đến sự cạnh tranh khốc liệt giữa các nhà làm phim. Dự báo chính xác về độ thành công của một bộ phim trước khi ra mắt có thể là điều kiện tiên quyết đối với quá trình ra quyết định có nên sản xuất một bộ phim, hay giúp các nhà làm phim đưa ra các định hướng tiếp thị và quảng bá.

Giải pháp tổng quan

Một trong những giải pháp có thể thực hiện để dự báo độ thành công của một bộ phim chính là dự đoán rating trước khi ra mắt bằng cách sử dụng các pre-released feature của các bộ phim bao gồm thông tin về diễn viên, đạo diễn, thể loại phim, ngày phát hành, và các yếu tố liên quan.

Để xây dựng được hệ thống dự đoán nói trên, cần sử dụng các phương pháp học máy có giám sát như Random Forest Regression, Support Vector Regression, hoặc các mô hình khác tùy thuộc vào yêu cầu và đặc điểm của dữ liệu.

Đầu tiên, thu thập dữ liệu về các bộ phim trên các nền tảng thông tin đáng tin cậy như IMDb, Rotten Tomatoes, MovieLens, … Sau khi đã có dữ liệu thô, có thể tiến hành các phương pháp feature engineering phù hợp để tiến hành huấn luyện mô hình dự đoán.

Tuy nhiên, dự đoán ratings của bộ phim chỉ dựa vào các pre-released features được liệt kê ở trên là một vấn đề khó khăn và không thể đảm bảo độ chính xác tuyệt đối. Do đó, các mô hình, hệ thống dự đoán này chỉ có thể đưa ra kết quả dựa mang tính chất tham khảo.

2. Thu thập và mô tả dữ liệu

2.1. Thu thập dữ liệu

2.1.1. Nguồn dữ liệu

IMDb là trang web hàng đầu về thông tin phim và ngành công nghiệp điện ảnh. Nó cung cấp một cơ sở dữ liệu phong phú về các bộ phim, diễn viên và đạo diễn. Người dùng có thể tìm kiếm và xem thông tin chi tiết về các bộ phim. IMDb cũng cho phép người dùng đánh giá và thảo luận với cộng đồng yêu thích điện ảnh. Đây là nguồn thông tin uy tín và hữu ích cho những người yêu thích điện ảnh.

2.1.2. Công cụ thu thập

BeautifulSoup là một thư viện Python mạnh mẽ được sử dụng để phân tích và trích xuất dữ liệu từ các trang web. Nó cung cấp các công cụ linh hoạt để truy cập và xử lý các thành phần HTML và XML trong một cách dễ dàng. Với BeautifulSoup, chúng ta có thể lấy thông tin từ các thẻ HTML, thuộc tính, văn bản và cấu trúc của một trang web. Chúng ta có thể tìm kiếm, lọc và trích xuất dữ liệu dựa trên các tiêu chí như lớp CSS, id, thẻ, v.v. Thư viện này giúp đơn giản hóa quá trình web scraping và phân tích dữ liệu từ trang web. Nó cung cấp một cú pháp dễ hiểu và mạnh mẽ để làm việc với HTML và XML.

2.1.3. Cách thức sử dụng công cụ

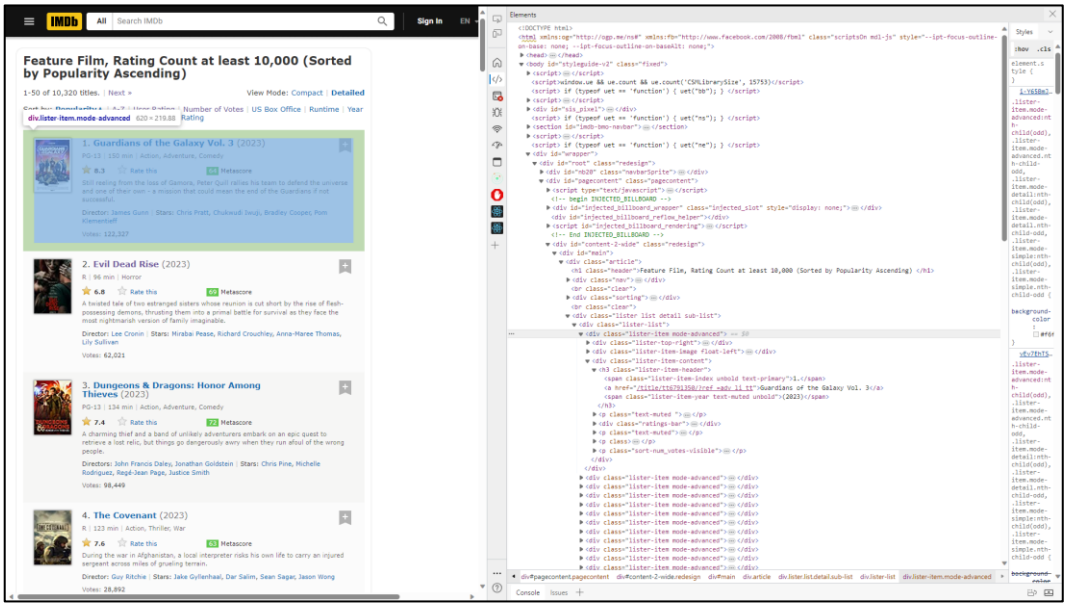
- Sử dụng thư viện requests để gửi yêu cầu GET đến trang web IMDb.
- Sử dụng thư viện BeautifulSoup để phân tích nội dung HTML trả về từ yêu cầu.
- Xác định các phần tử trên trang web chứa thông tin cần thu thập (ví dụ: tiêu đề phim, năm phát hành, đánh giá IMDb, số phiếu bầu, thể loại, thời lượng, v.v.).
- Trích xuất thông tin từ các phần tử và lưu trữ vào các biến hoặc mảng.
- Lặp lại quá trình trên các trang khác nhau để thu thập thông tin từ nhiều trang.

2.1.4. Đầu vào và đầu ra của quá trình thu thập

- Đầu vào: URL của trang web IMDb gồm các tham số truy vấn để tìm kiếm, lọc và sắp xếp danh sách phim theo yêu cầu.
- Đầu ra: Dữ liệu thu thập được từ trang web IMDb, bao gồm thông tin về các bộ phim như title, releaded, imdb score, genre, runtime, votes, rating, directors, stars, v.v. Dữ liệu này được lưu trữ trong các cấu trúc dữ liệu như mảng, DataFrame sau đó được lưu vào file (ví dụ: CSV, Excel) để sử dụng và phân tích sau này.

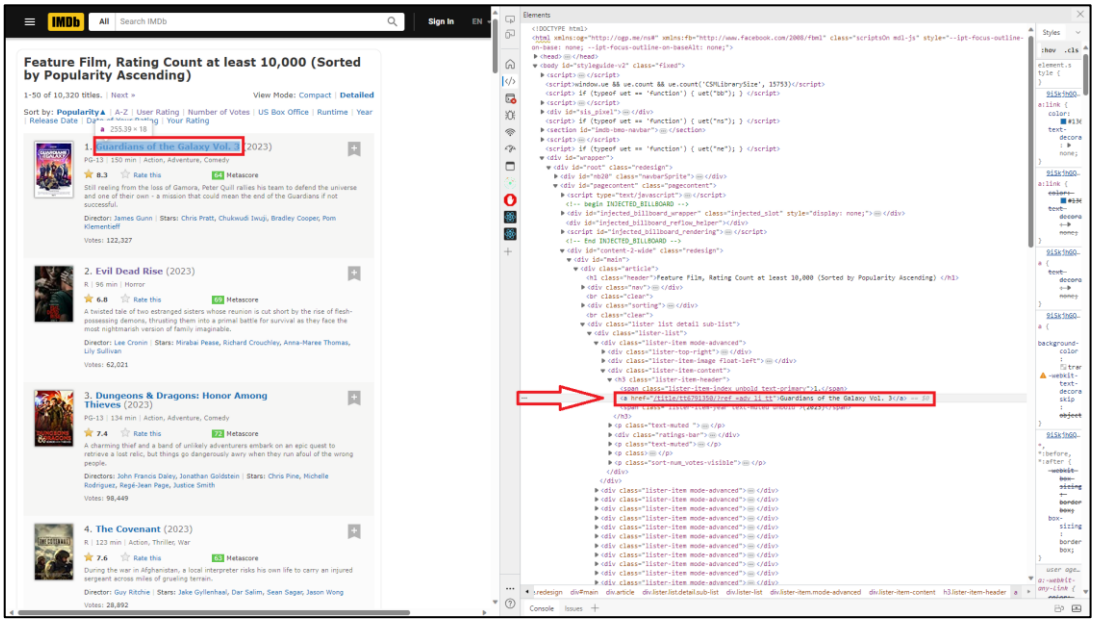
2.1.5. Ví dụ minh hoạ

- Thực hiện hàm get() của thư viện requests ta để lấy được nội dung HTML của trang. Tiếp đó, sử dụng hàm BeautifulSoup() để phân tích trang HTML.
- Sau đó thực hiện hàm find_all() của thư viện BeautifulSoup, ta sẽ lấy được tất các thẻ ‘div’ có class là ‘lister-item mode-advanced’. Đây chính là các container chứa thông tin của phim.



Hình 1. Mô tả cách lấy ra container chứa thông tin của phim.

- Ở ví dụ này, ta sẽ lấy đặc trưng "title" của phim:
  - Xác định vị trí của nó trong DOM của trang.
  - Sử dụng find() để truy vấn đến thẻ đó.
  - Sử dụng thuộc tính text để lấy ra nội dung của thẻ.



Hình 2. Mô tả cách lấy ra đặc trưng 'title' của phim.

- Thực hiện tương tự bước trên với các thông tin khác bao gồm tiêu đề, năm phát hành, đánh giá IMDb, số phiếu bầu, thể loại, thời lượng, công ty sản xuất, quốc gia, ngân sách, doanh thu, đạo diễn, biên kịch và diễn viên chính. Dữ liệu thu thập được lưu trữ trong một DataFrame và sau đó được xuất ra file CSV để sử dụng và phân tích.

2.2. Mô tả dữ liệu

2.2.1. Thống kê tổng quan về tập dữ liệu

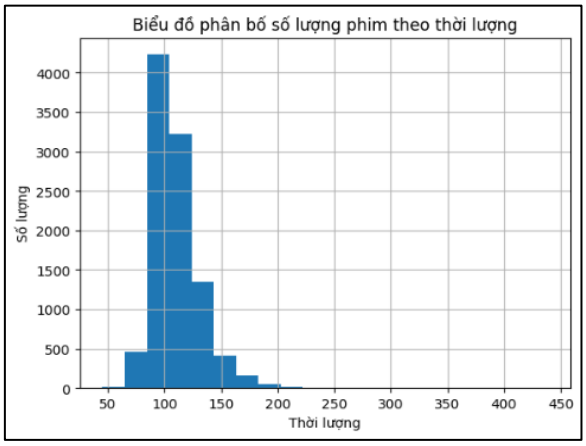
STT	Đặc trưng	Mô tả	Kiểu dữ liệu
1	title	Tên của phim	object
2	year	Năm phim ra mắt	int64
3	genre	Các thể loại của phim	object
4	runtime	Thời lượng của phim	int64
5	imdb	Điểm số IMDb	float64
6	votes	Số lượt bình chọn trên trang IMDb của phim	int64
7	released	Ngày ra mắt của phim	object
8	budget	Kinh phí của phim	float64
9	companies	Các công ty sản xuất	object
10	gross	Doanh thu của phim	float64
11	diretors	Các đạo diễn tham gia	object
12	writers	Các biên kịch tham gia	object
13	stars	Các diễn viên chính tham gia đóng phim	object
14	origins	Những nơi công chiếu phim đầu tiên	object
15	rating	Giới hạn độ tuổi(R, PG-13,…)	object

Bảng 1. Thống kê các đặc trưng trong tập dữ liệu.

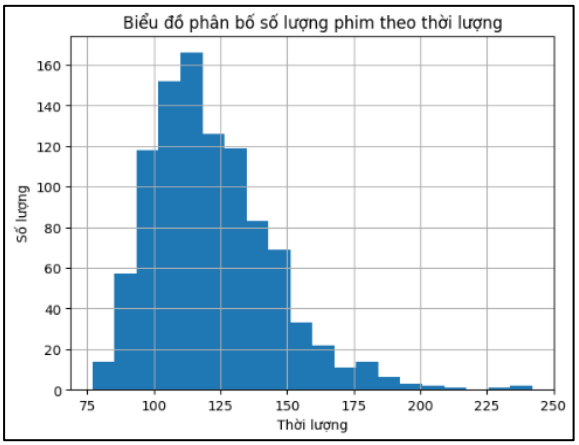
	BIG DS	SMALL DS
Tổng dữ liệu (records)	10000	1000
Dữ liệu trống trong đặc trựng “rating”	277	4
Dữ liệu trống trong đặc trựng “budgets”	3460	128
Dữ liệu trống trong đặc trựng “gross”	1225	39

Bảng 2. Thống kê tổng quan tập dữ liệu.

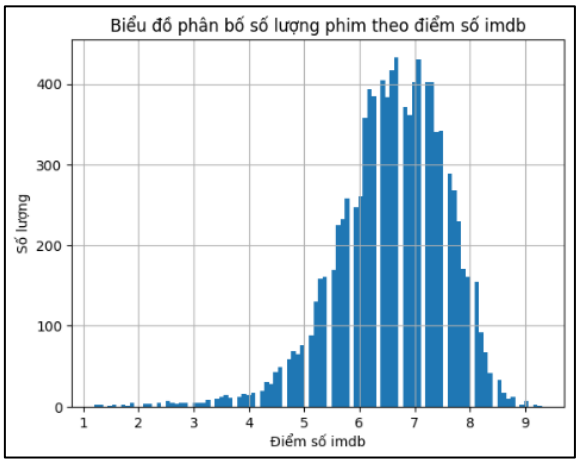
2.2.2. Mô tả trực quan một số đặc trưng đáng chú ý



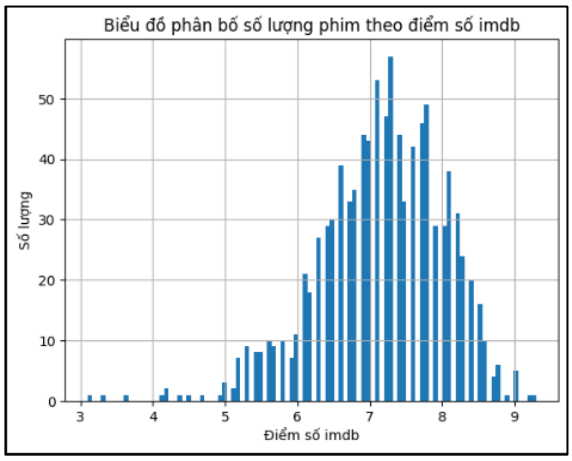
Hình 4. Phân bố số lượng phim theo thời lượng của Big DS



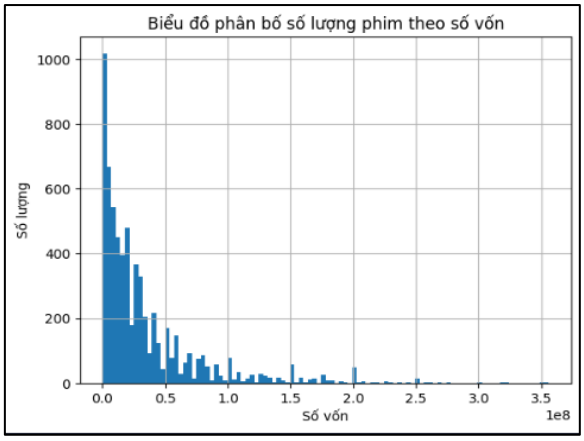
Hình 3. Phân bố số lượng phim theo thời lượng của Small DS



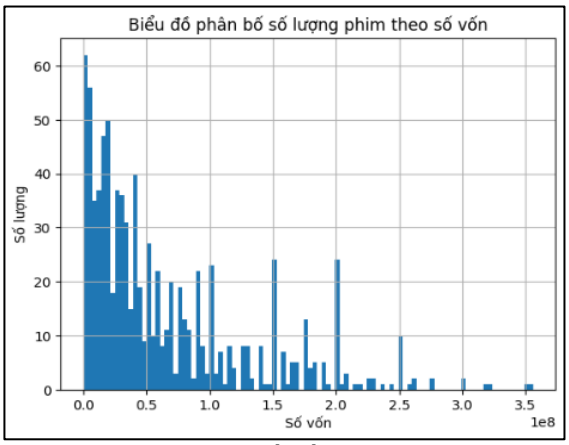
Hình 6. Phân bố số lượng phim theo điểm số imdb của Big DS



Hình 5. Phân bố số lượng phim theo điểm số imdb của Small DS



Hình 8. Phân bố số lượng phim theo số vốn đầu tư của Small DS.



Hình 7. Phân bố số lượng phim theo số vốn đầu tư của Small DS.

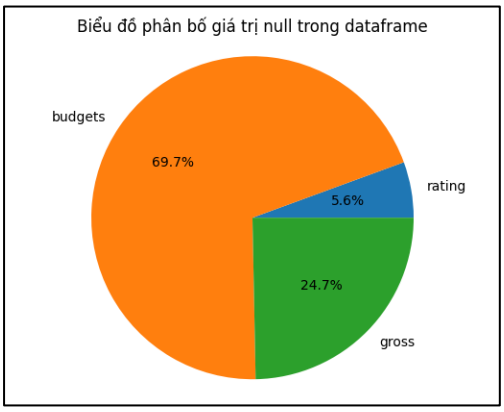
Từ các đồ thị trên, ta cũng rút ra được một số nhận xét sau:

- Thời lượng phim 80 – 140 phút có số lượng cao. Điều này liên quan đến thói quen khi đi xem phim của mọi người, các phim quá dài làm cho người xem chán nản, phim quá ngắn thì không đủ xây dựng nhân vật và tình tiết.
- Điểm IMDB có phổ điểm rộng 1 – 9, nhưng tập rất nhiều ở mức điểm 6 – 8. Đây là mức điểm khá tốt đối với một bộ phim.
- Kinh phí bỏ ra cho một bộ phim cao nhất gần 350.000.000 dollar, đa số thì bé hơn 50.000.000 dollar.

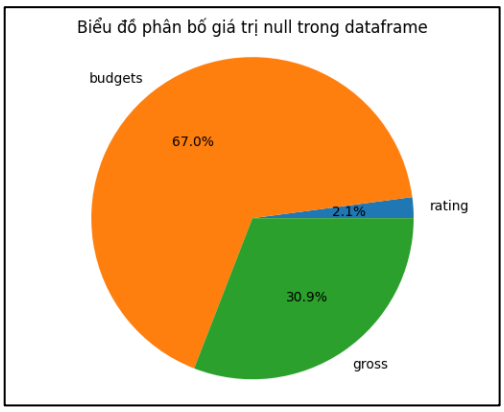
3. Trích xuất đặc trưng

3.1. Làm sạch dữ liệu

3.1.1. Xử lý dữ liệu trống



Hình 9. Phân bố giá trị null của Big DS.



Hình 10. Phân bố giá trị null của Small DS.

Tập dữ liệu chứa các trường chứ dữ liệu trống đó là budgets, rating và gross. Tuy nhiên, việc xử lý các trường này bằng các kỹ thuật Mean/Median/Mode/... sẽ không đảm bảo được tính chính xác của dữ liệu. Thay vào đó ta sẽ xóa các dòng chứa giá trị trống đi để đảm bảo được tính chính xác của dữ liệu.

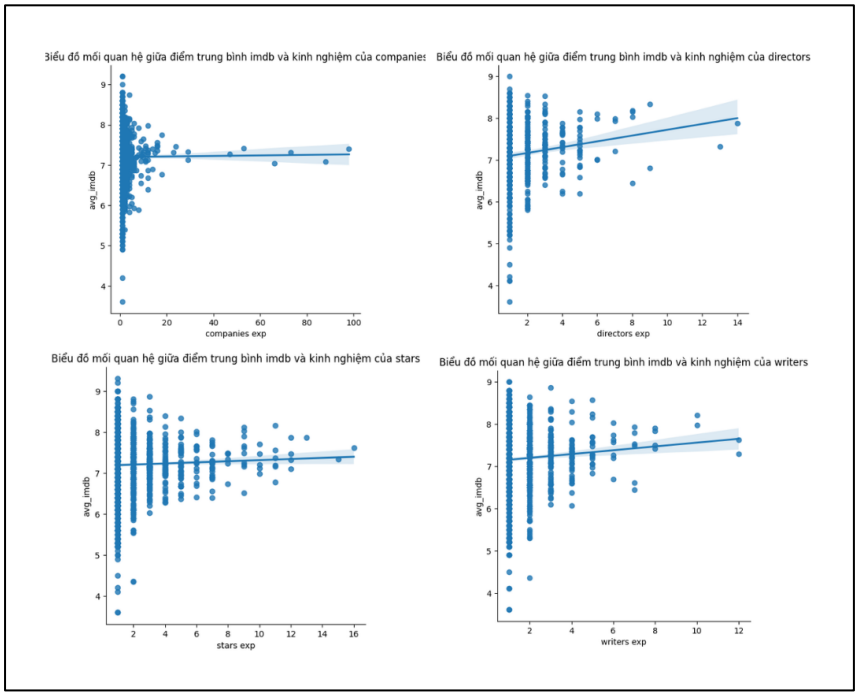
	BIG DS	SMALL DS
Tổng dữ liệu trước khi loại bỏ giá trị trống (records)	10000	1000
Tổng dữ liệu sau khi loại bỏ giá trị trống (records)	6053	851
Phần trăm mất mát dữ liệu sau khi loại bỏ các giá trị trống.	39.5%	14.9%

Bảng 3. Thống kê dữ liệu trước và sau khi loại bỏ giá trị trống.

3.1.2. Xử lý Companies, Writers, Stars, Directors

Có thể tạo thêm đặc trưng mới từ danh sách các công ty. Đặc trưng mới này sẽ tương ứng với kinh nghiệm của công ty đó, trích xuất bằng cách thống kê số lượng phim mà công ty đã tham gia sản xuất trong dataset. Một lưu ý chính là nếu phim có nhiều hơn một công ty tham gia sản xuất, ta sẽ chỉ xét công ty có kinh nghiệm lớn nhất. Áp dụng tương tự với kịch bản, diễn viên, đạo diễn.

Lúc này ta có thể tạo ra mối quan hệ của điểm imdb và kinh nghiệm của companies/writers/stars/directors. Giả sử, một đạo diễn có kinh nghiệm thì điểm imdb của phim có ảnh hưởng so với một đạo diễn có ít kinh nghiệm hơn không?



Hình 11. Biểu đồ mối quan hệ giữa điểm imdb và kinh nghiệm của companies/writers/stars/writers của Small DS.

Từ các đồ thị trên, ta kết luận được nếu creators(directors/writers/stars/companies) của một bộ phim có càng nhiều kinh nghiệm thì điểm imdb của phim sẽ càng lớn. Điều này cho thấy rằng kinh nghiệm của creators cũng sẽ ảnh hưởng đến điểm imdb của phim.

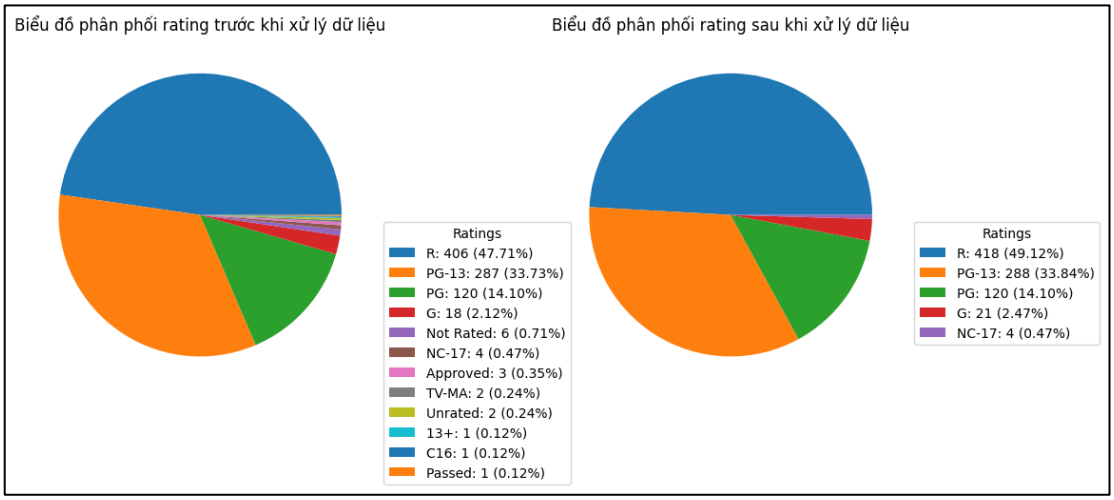
3.1.3. Xử lý Rating, Origins

Có rất nhiều rating và origins đã được thu thập trong dataset.

Trong bài toán này ta sẽ giữ lại 5 rating phổ biến nhất là R, PG-13, PG, G, NC-17. Ta tiến hành thay thế tất cả các rating trong dataset bằng 5 rating phổ biến như sau:

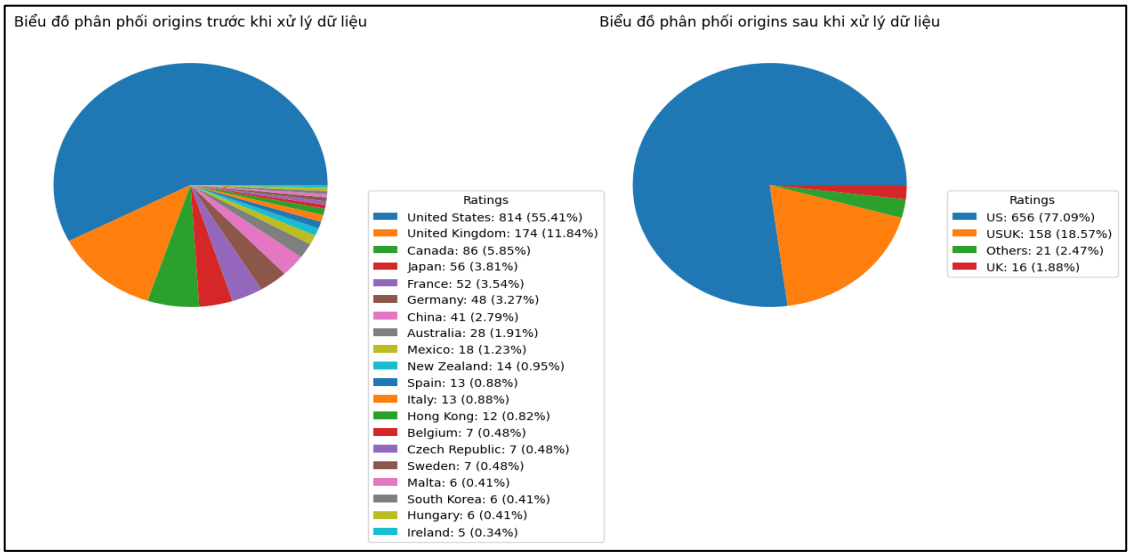
13+ = TV-14 = **PG-13**, 16+ = C16 = TV-MA = Not Rated = Passed = Unrated = **R**.

M = M/PG = GP = TV-PG = **PG**, X = **NC-17**, Approved = TV-G = **G**.



Hình 12. Biểu đồ phân phối rating trước và sau khi xử lý dữ liệu của Small DS.

Orgins của các bộ phim chiếm đa số ở United States và United Kingdom, còn lại sẽ nằm rải rác ở các quốc gia khác. Do đó ta sẽ chia lại lớp cho trường Orgins như sau: US (United State), UK (United Kingdom), USUK (phim có origin gồm cả United State và United Kingdom) và Others.



Hình 13. Biểu đồ phân phối origins trước và sau khi xử lý dữ liệu của Small DS.

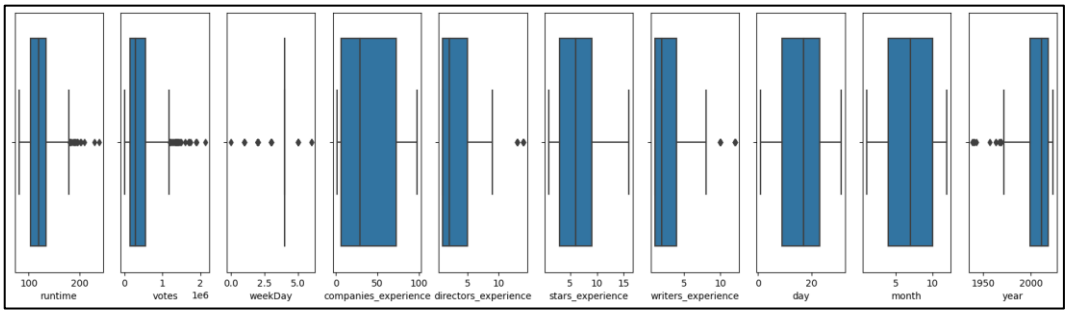
3.1.4. Xử lý Gernes

Gernes bao gồm nhiều thể loại của một bộ phim. Ví dụ: phim chỉ có thể loại Comedy sẽ khác so với phim có cả thể loại Action và Comedy. Xử lý đặc trưng này bằng cách biến đổi mỗi danh mục gernes thành một con số riêng biệt, sử dụng Label Encoder.

3.1.5. Xử lý Releaseds

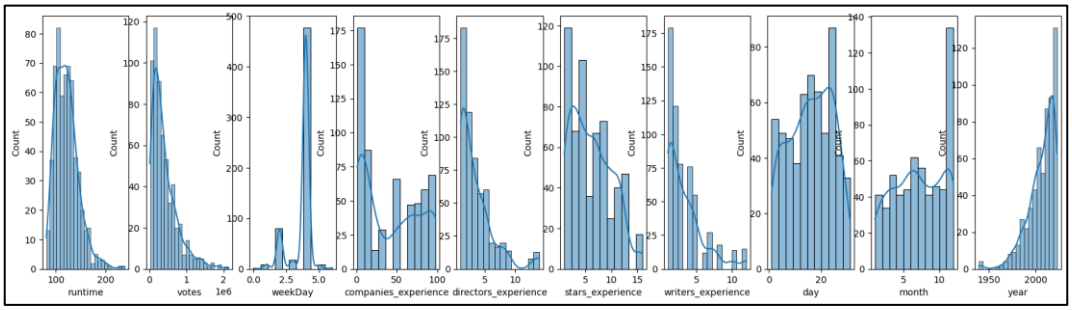
Ta sẽ tách releaseds thành 4 cột mới là: day, month, year, weekDay, trong đó: day, month, year sẽ là day, month, year của ngày phát hành bộ phim; weekDay là ngày trong tuần của ngày phát hành bộ phim - từ thứ Hai đến Chủ Nhật.

3.2. Xử lý ngoại lệ



Hình 14. Đồ thị boxplot của các trường dữ liệu trong Small DS.

Từ đồ thị boxplot đã nêu ở trên, ta có thể nhận thấy 2 trường có nhiều ngoại lệ nhất đó chính là runtime và votes, nên ta sẽ tiến hành xử lý ngoại lệ cho 2 trường này.



Hình 15. Histogram của các trường dữ liệu trong Small DS.

Trong khi runtime có kiểu phân bố gần giống với phân phối chuẩn gauss thì votes có kiểu phân bố lệch, ta sẽ thực hiện xử lý ngoại lệ theo 2 cách khác nhau đó là "gauss" và "skew".

3.3. Chuẩn hóa dữ liệu

Để cải thiện độ chính xác, tốc độ thực thi của mô hình, cũng như giúp mô hình học máy dễ dàng hội tụ và tạo kết quả tốt hơn, dữ liệu phải cần được chuẩn hóa. Trong các phương pháp chuẩn hóa dữ liệu, phương pháp Min Max Scaler được sử dụng phổ biến và áp dụng hiệu quả với dữ liệu gốc của bài toán – dữ liệu có phân phối không chuẩn. Phương pháp Min Max Scaler chuyển dữ liệu gốc vào một phạm vi cụ thể như [0, 1] bởi công thức:

$$x_{std} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$$x_{scaled} = x_{std} * (max - min) + min$$

Trong đó: min = 0 và max = 1.

3.4. Lựa chọn đặc trưng

Thuật toán RFE (Recursive Feature Elimination) được sử dụng để lựa chọn đặc trưng. RFE hoạt động dựa vào một external estimator (có thể là các model như SVR, RFR được sử

dụng trong tiểu luận này). Estimator này sẽ được huấn luyện trên tập đặc trưng ban đầu, sau đó tầm quan trọng các các đặc trưng sẽ được tính toán để loại bỏ đặc trưng ít quan trọng nhất. Quá trình này sẽ được đệ quy cho đến khi nào tập đặc trưng đưa vào ban dầu đạt đến số lượng mong muốn. Kết quả mô hình dự đoán trước và sau sử dụng RFE để lựa chọn đặc trưng được trình bày cụ thể tại phần ***4.3.3***.

Thư viện scikit-learn cung cấp lớp sklearn.feature_selection.**RFE** để có thể cài đặt thuật toán này. Một vài tham số sử dụng:

Tham số	Ý nghĩa
estimator	Mô hình học máy có giám sát, cung cấp phương thức "fit" để tính toán "feature importance".
n_features_to_select	Số lượng đặc trưng cần lựa chọn.
step	Là giá trị lớn hơn 1, biểu thị số lượng đặc trưng muốn loại bỏ sau mỗi lần lặp.

Bảng 4. Tham số sử sụng trong RFE

4. Mô hình hóa dữ liệu

Hai trong số các mô hình phổ biến được sử dụng trong các bài toán dự đoán hồi quy bao gồm: Random Forest Regression và Support Vector Regression.

4.1. Random Forests

Random Forest là bộ dự đoán tổng hợp (meta-estimator) sử dụng nhiều cây quyết định phân loại khác nhau trên nhiều tập dữ liệu con khác nhau của tập dữ liệu ban đầu. Mô hình sử dụng phương pháp trung bình để cải thiện độ chính xác dự đoán và kiểm soát hiện tượng over-fitting. Mô hình học máy Random Forest được sử dụng cho cả bài toán phân lớp và bài toán hồi quy. Tiểu luận này sử dụng lớp ***RandomForestRegression*** từ module ***sklearn.ensemble*** để xây dựng hệ thống dự đoán hồi quy. Một vài tham số sử dụng huấn luyện mô hình:

Tham số	Ý nghĩa
n_estimators	Số lượng cây trong mô hình forest.
max_depth	Độ sâu tối đa của cây.
min_samples_split	Số lượng mẫu tối thiểu để chia một nút trong cây.
min_samples_leaf	Số lượng mẫu tối thiểu trong mỗi lá của cây.
max_features	Số lượng đặc trưng được xem xét khi tìm kiếm phân chia tốt nhất.
random_state	Giá trị đảm bảo kết quả được tái tạo như nhau trong quá trình huấn luyện mô hình.

Bảng 5. Tham số sử dụng huấn luyện mô hình Random Forest Regression

4.2. Support Vector Machines

SVMs là một trong những mô hình học máy có giám sát được sử dụng trong các bài toán phân lớp, hồi quy và dự đoán ngoại lệ. Đối với bài toán hồi quy, có thể sử dụng Support Vector Regression (SVR) với ý tưởng tương tự như bài toán phân lớp. Mục tiêu của SVR là tìm ra hàm xấp xỉ mối quan hệ giữa các biến đầu vào với một biến đầu ra mang giá trị liên tục, với lỗi dự đoán tối thiểu. Không như bài toán phân lớp, SVR tìm kiếm hyperlane phù hợp với điểm dữ liệu trong không gian liên tục.

Thư viện scikit.learn cung cấp lớp sklearn.svm.SVR để có thể xây dựng mô hình dự đoán hồi quy cho bài toán Predicting move ratings. Một vài tham số sử dụng huấn luyện:

Tham số	Ý nghĩa
kernel	Chỉ định loại kernel để tính toán trước kernel matrix.
C	Tham số tổng quát hóa (giúp mô hình tránh overfitting) với penalty là l2 squared.
epsilon	Tham số xác định phạm vi chấp nhận được cho sai số của các mẫu huấn luyện, quyết định mức độ linh hoạt của đường biên hỗ trợ (support boundary) xung quanh các điểm dữ liệu huấn luyện.

Bảng 6. Tham số sử dụng huấn luyện mô hình Support Vector Regression

4.3. Thực nghiệm và kết quả

4.3.1. Dữ liệu

Quá trình thực nghiệm sử dụng hai dataset được crawl cùng một nguồn với các kích thước khác nhau được mô tả chi tiết tại mục ***2.2***. Hai dataset này được chia thành các tập Train/Validation/Test với cùng tỉ lệ như nhau là 70/15/15.

4.3.2. Metrics đánh giá

• ***Mean absolute error (MAE):*** đo độ lớn trung bình của các lỗi trong một tập hợp các dự đoán mà không cần xem xét hướng của chúng. MAE ít bị ảnh hưởng bởi các giá trị lỗi lớn, giá trị MAE càng thấp thì mô hình dự đoán càng chính xác.

$$MAE = \frac{\Sigma_{i=1}^{n}|y_i - x_i|}{n}$$

Trong đó, n là tổng số điểm dữ liệu, y_i là giá trị dự đoán, x_i là giá trị thực.

• ***Root mean squared error (RMSE):*** đo độ lớn trung bình của các các lỗi, là khoảng cách trung bình từ điểm dữ liệu đến *fitted line*. RMSE bị ảnh hưởng nhiều bởi các giá trị lỗi lớn. RMSE có giá trị càng thấp thì mô hình dự đoán càng tốt.

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

Trong đó, N là tổng số điểm dữ liệu, x_i là giá trị thực, $\hat{x}_i$ là giá trị dự đoán.

• ***Coeffificient of determination (R^2):*** R2 đo lường mức độ giải thích của mô hình đối với biến phụ thuộc. Nó biểu thị tỷ lệ phần trăm của sự biến động của biến phụ thuộc được giải thích bởi mô hình. R2 càng gần 1, mô hình càng tốt và có khả năng dự đoán tốt hơn.

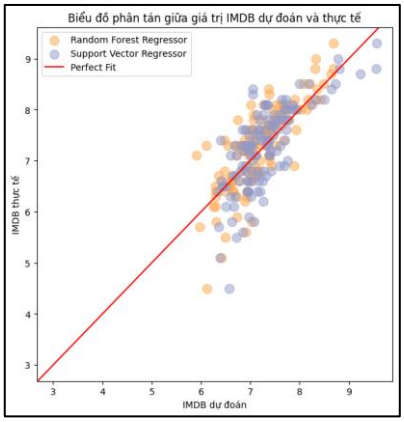
$$R^2 = 1 - \frac{\Sigma_{i=1}^{n}(\hat{y}_i - y_i)^2}{\Sigma_{i=1}^{n}(y_i - \overline{y_i})^2}$$

Trong đó, y_i là giá trị thực tế, $\hat{y}_i$ là giá trị dự đoán, $\overline{y_i}$ là trung bình của các giá trị.

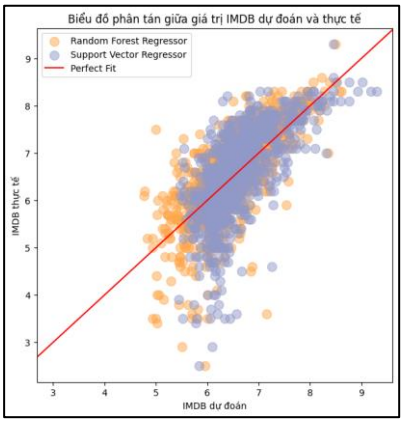
4.3.3. Kết quả

Kết quả mô hình dự đoán đối với tập dataset với kích thước nhỏ (1000 mẫu) cho thấy kết quả tốt hơn so với tập kích thước lớn (10000 mẫu). Điều này xảy ra bởi tập dữ liệu nhỏ có dữ liệu không đa dạng bằng tập lớn, do đó mô hình có thể dễ dàng fit được dữ liệu và dự đoán tốt trên tập validation và cả tập test (150 mẫu).

Trước lựa chọn đặc trưng:



Hình 17. Trực quan giá trị IMDB dự đoán và thực tế trên Small Dataset

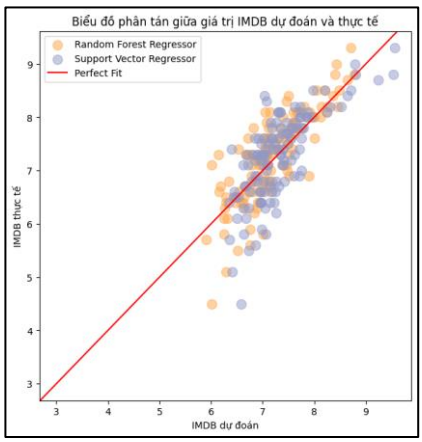


Hình 16. Trực quan giá trị IMDB rating dự đoán và thực tế trên Big Dataset

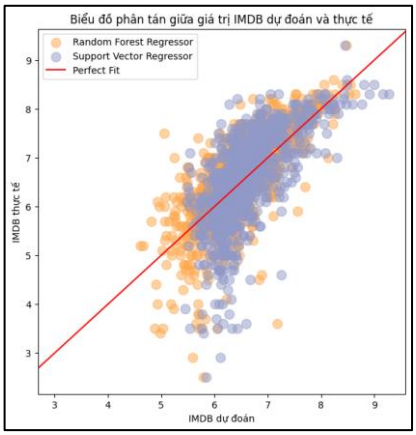
Metrics	Small Dataset (1000 mẫu)		Big Dataset (10000 mẫu)	
	RF	SVM	RF	SVM
MAE	0.39	0.42	0.47	0.52
RMSE	0.51	0.55	0.64	0.70
R^2	0.61	0.55	0.55	0.45

Bảng 7. Kết quả các metrics đánh giá trên tập Validation trước lựa chọn đặc trưng

Sau lựa chọn đặc trưng:



Hình 18. Trực quan giá trị IMDB rating dự đoán và thực tế trên Small Dataset

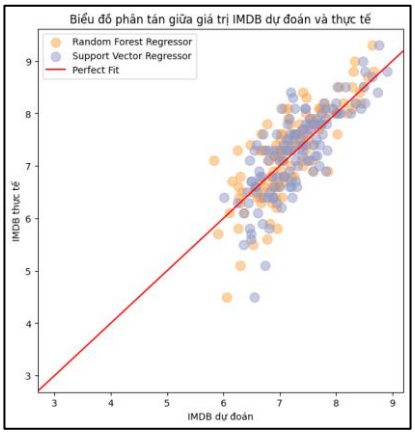


Hình 19. Trực quan giá trị IMDB rating dự đoán và thực tế trên Big Dataset

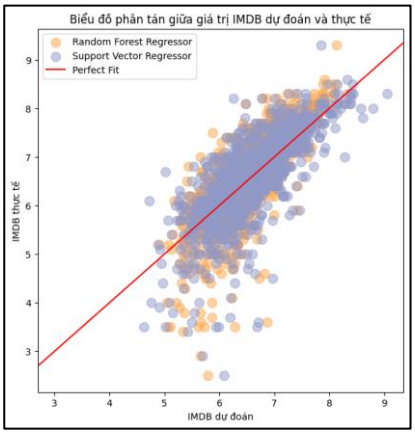
Metrics	Small Dataset (1000 mẫu)		Big Dataset (10000 mẫu)	
	RF	SVM	RF	SVM
MAE	0.39	0.42	0.47	0.52
RMSE	0.51	0.55	0.63	0.70
R^2	0.61	0.55	0.56	0.45

Bảng 8. Kết quả các metrics đánh giá trên tập Validation sau lựa chọn đặc trưng

Sau tối ưu tham số



Hình 21. Trực quan giá trị IMDB rating dự đoán và thực tế trên Small Dataset



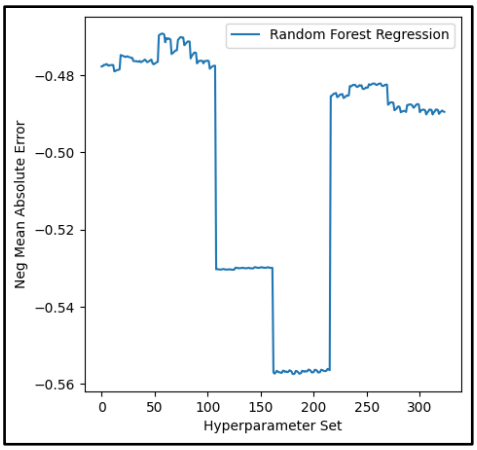
Hình 20. Trực quan giá trị IMDB rating dự đoán và thực tế trên Big Dataset

Metrics	Small Dataset (1000 mẫu)		Big Dataset (10000 mẫu)	
	RF	SVM	RF	SVM
MAE	0.38	0.38	0.46	0.46
RMSE	0.49	0.49	0.61	0.63
R^2	0.63	0.64	0.58	0.56

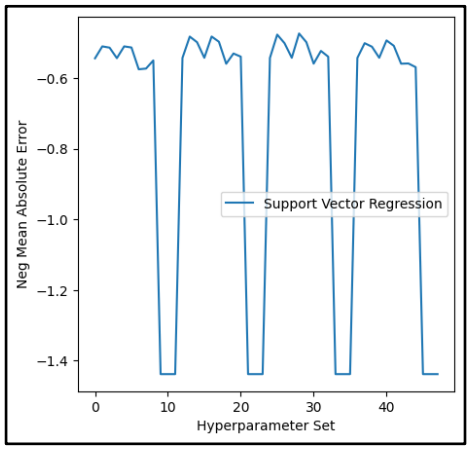
Bảng 9. Kết quả các metrics đánh giá trên tập Validation sau tối ưu tham số

Nhìn vào các đồ thị phân tán giữa kết quả IMDB rating dự đoán và IMDB rating thực tế, có thể dễ dàng nhận thấy cả hai mô hình (SVR và RFR) đều chưa dự đoán tốt các IMDB

rating có giá trị thấp (<5). Đều này có thể lí giải bởi phân bố dữ liệu IMDB rating từ dataset crawl về (mục ***2.2***), dữ liệu các phim có IMDB rating thấp vẫn còn ít, chiếm < 1%.



Hình 22. Trực quan quá trình tìm siêu tham số cho mô hình RFR



Hình 23. Trực quan quá trình tìm siêu tham số cho mô hình SVR

Từ các bảng kết quả metrics đánh giá sau lựa chọn tham số tối ưu cho thấy RF có khả năng dự đoán tương tự với SVM, với các giá trị MAE, RMSE chênh lệch nhau ± 0.01, và R^2 chênh lệch nhau ± 0.02 (với tập Big Dataset); kết quả tương tự đối với tập Small Dataset. Bằng trực quan quá trình tìm siêu tham số cho các mô hình, có thể nhận thấy SVR phụ thuộc nhiều vào cấu hình chính xác, cụ thể là các tham số tối ưu cần phù hợp, để có thể mang lại hiệu quả tốt như RFR (MAE giảm 0.067, RMSE giảm 0.075 và R^2 tăng 0.111 sau khi tối ưu tham số với Big Dataset). So sánh trực quan các kết quả của toàn bộ quá trình huấn luyện tại hai bảng kết quả sau:

	Model	Step	MAE	RMSE	R2	MAE Improvement	RMSE Improvement	R2 Improvement
0	Random Forest	All Features	0.390	0.509	0.609	-	-	-
1	SVM	All Features	0.412	0.538	0.564	-	-	-
2	Random Forest	Selected Features	0.392	0.507	0.612	0.001	-0.002	0.003
3	SVM	Selected Features	0.420	0.548	0.548	0.007	0.009	-0.015
4	Random Forest	Tuned Hyperparams	0.379	0.493	0.633	-0.013	-0.014	0.021
5	SVM	Tuned Hyperparams	0.380	0.492	0.635	-0.04	-0.055	0.087

Hình 24. Tổng hợp kết quả metrics đánh giá trên Small Dataset

	Model	Step	MAE	RMSE	R2	MAE Improvement	RMSE Improvement	R2 Improvement
0	Random Forest	All Features	0.472	0.639	0.545	-	-	-
1	SVM	All Features	0.523	0.702	0.451	-	-	-
2	Random Forest	Selected Features	0.470	0.636	0.550	-0.002	-0.003	0.005
3	SVM	Selected Features	0.523	0.702	0.451	0.0	0.0	-0.0
4	Random Forest	Tuned Hyperparams	0.458	0.614	0.580	-0.012	-0.022	0.031
5	SVM	Tuned Hyperparams	0.456	0.627	0.562	-0.067	-0.075	0.111

Hình 25. Tổng hợp kết quả các metrics đánh giá trên Big Dataset

Kết quả trên tập Test:

Metrics	Small Dataset (1000 mẫu)		Big Dataset (10000 mẫu)	
	RF	SVM	RF	SVM
MAE	0.39	0.37	0.47	0.46
RMSE	0.53	0.51	0.64	0.65
R^2	0.60	0.63	0.61	0.59

Bảng 10. Kết quả các metrics đánh giá trên tập Test

Với tập dữ liệu hoàn toàn mới, cả hai mô hình đều cho ra kết quả dự đoán khá tốt và tương đồng nhau, điều này cho thấy mô hình không gặp các vấn đề như overfitting và underfitting. Mô hình khớp được ~60% dữ liệu và có khả năng dự đoán tốt IMDB rating của bộ phim (thông qua các giá trị MAE, RMSE, R^2 được trình bày ở bảng trên).

5. Kết luận

Tiểu luận này đã trình bày phương pháp cho bài toán dự đoán IMDB rating của các bộ phim trước khi ra mắt, dựa vào các pre-released feature của các bộ phim. Các giải pháp bao gồm cách thu thập dữ liệu, xử lý dữ liệu và các phương pháp, mô hình học máy sử dụng. Hai mô hình Random Forest Regression và Support Vector Regression được triển khai mang lại kết quả dự đoán khá tốt, trong đó RFR mang lại kết quả tốt đối với tập dữ liệu Validation, SVR mang lại kết quả tốt hơn đối với tập dữ liệu Test hoàn toàn mới.

Tuy nhiên, với sự đa dạng dữ liệu chưa đồng nhất, cụ thể là phân bố điểm IMDB rating của tập dữ liệu dùng để huấn luyện chưa đồng đều, có rất ít các bộ phim có điểm IMDB rating dưới 5 khiến mô hình cho kết quả dự đoán với sai số lớn. Trong tương lai, để cải thiện độ chính xác của mô hình, có thể thu thập thêm các dữ liệu phim có rating dưới 5, khiến dữ liệu đa dạng hơn.

6. Tài liệu tham khảo

[Bachelor Degree Project in Informatics G2E, ECTS Spring term 2015] **Karl Persson**, PREDICTING MOVIE RATINGS - A comparative study on random forests and support vector machines

[1-12018] **Yueming Zhang**, Predict IMDB score with data mining algorithms

[March 2023] **Jyoti Tripathi, Sunita Tiwari, Anu Saini, Sunita Kumari**, Prediction of movie success based on machine learning and twitter sentiment analysis using internet movie database data.

[Stockholm, Sverige 2017] **Pojan Shahrivar, Carl Jernbacker**, Predicting movie success using machine learning techniques.