TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN CUỐI KỲ

HỌC PHẦN: KHOA HỌC DỮ LIỆU

ĐỀ TÀI: DỰ ĐOÁN MỨC DOANH THU PHIM

HỌ VÀ TÊN SINH VIÊN	LỚP HỌC PHẦN	ĐIỂM BẢO VỆ
NGUYỄN TRẦN THẢO VY	20Nh91	
NGUYỄN ĐỨC QUỐC	20Nh91	
LÊ HOÀNG	20Nh91	

ĐÀ NẴNG, 05/2023

TÓM TẮT

Dựa vào dữ liệu rất lớn của các bộ phim được tổng hợp trên website của The number, chúng em đã lựa chọn đề tài “dự đoán mức doanh thu của phim”. Sau khi nghiên cứu, nhóm đã quyết định dùng thư viện request và beautifulsoup để cào data; Label Encoder, Standard Scaling và Outliers Handling để trích xuất đặc trưng; hai mô hình là Logistic Regression và Random Forest để dự đoán doanh thu mức doanh thu phim; sau cùng là sử dụng Accuary, Precision, Recall, F1-Score để đánh giá các mô hình. Kết quả là nhóm đã cào được Data từ web, trích xuất đặc trưng cũng như dùng hai mô hình để đánh giá mức doanh thu phim.

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá theo 3 mức (Đã hoàn thành/Chưa hoàn thành/Không triển khai)
Nguyễn Đức Quốc	- Làm sach dữ liệu, xử lý dữ liệu trống - Xử lý ngoại lệ - Mã hóa dữ liệu - Lựa chọn đặc trưng	- Đã hoàn thành - Đã hoàn thành - Đã hoàn thành - Đã hoàn thành
Nguyễn Trần Thảo Vy	- Thu thập dữ liệu - Thống kê mô tả trực quan về dữ liệu - Dán nhãn dữ liệu	- Đã hoàn thành - Đã hoàn thành - Đã hoàn thành
Lê Hoàng	- Mô hình hóa dữ liệu - Tìm bộ siêu tham số cho mô hình - Chuẩn hóa dữ liệu - Đánh giá hiệu quả mô hình	- Đã hoàn thành - Đã hoàn thành - Đã hoàn thành - Đã hoàn thành

MỤC LỤC

DANH MỤC HÌNH ẢNH

1. Giới thiệu

1.1 Các vấn đề cần giải quyết

- Làm sao để cào được dữ liệu từ web.

- Cách để trích xuất dữ liệu

- Phương pháp để dự đoán doanh thu phim và độ chính xác của phương pháp đó

1.2 Giải pháp

- Sử dụng thư viện request, beautifulsoup phục vụ cho việc cào dữ liệu.

- Quan sát bằng mắt và chọn lọc, sau đó cào các đặc trưng đã chọn tương ứng với các thẻ HTML.

- Sử dụng hai mô hình học máy là RandomForest và Linear Regression.

2. Thu thập và mô tả dữ liệu

2.1. Thu thập dữ liệu

2.1.1. Nguồn dữ liệu

Dữ liệu được thu thập từ trang web thống kê, đánh giá về ngành công nghiệp điện ảnh và truyền hình thế giới: https://www.the-numbers.com/

2.1.2. Công cụ thu thập

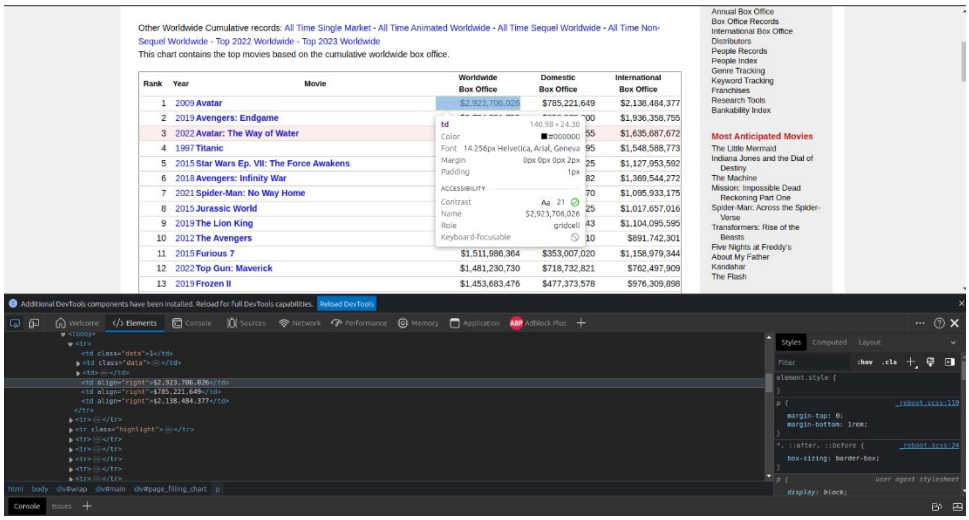
- IDE: Visual Studio Code

- Ngôn ngữ: Python

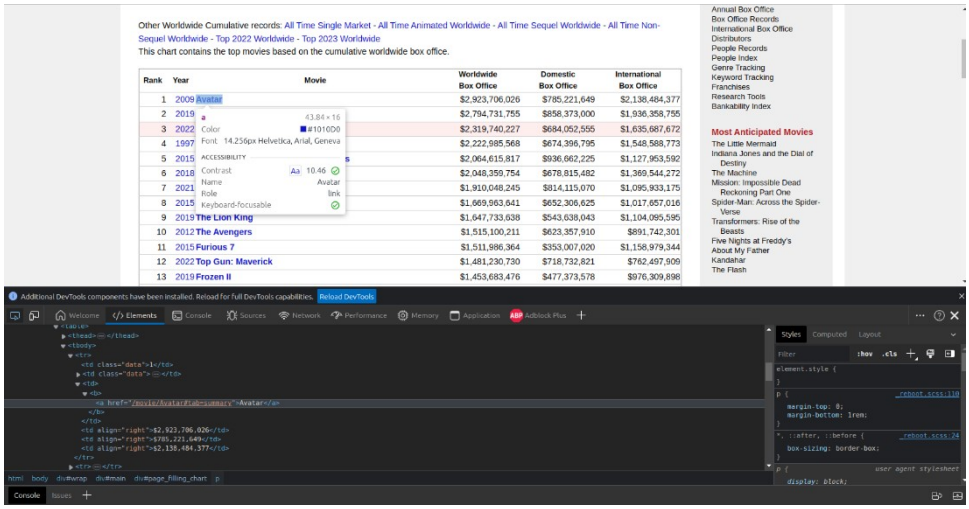
- Thư viện: beautifulsoup4, request

2.1.3. Cách thức thu thập

Bước 1: Lấy tất cả doanh thu và link thông tin chi tiết của các bộ phim

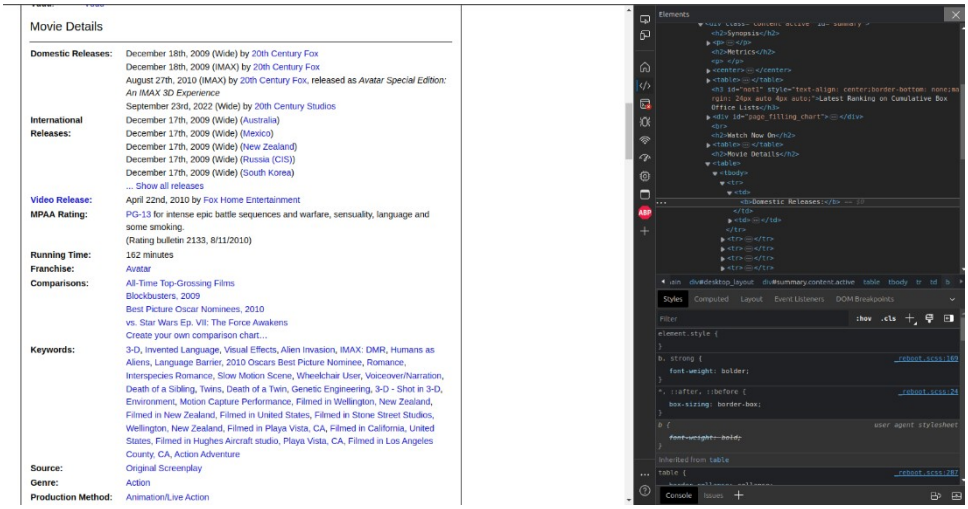


Hình 1. Thẻ chứa doanh thu toàn cầu của một bộ phim



Hình 2. Thẻ chứa link đến thông tin chi tiết của một bộ phim

Bước 2: Phân tích cấu trúc trang web của từng bộ phim



Hình 3. Phân tích cấu trúc HTML của một bộ phim

- Tìm kiếm các thẻ chứa thông tin cần thiết của bộ phim.

Bước 3: Trích xuất thông tin

- Sử dụng thư viện request để get trang web.

- Đọc HTML của trang web thông qua hàm của thư viên beautifulsoup.

Bước 4: Chuyển list dữ liệu sang file csv

- Input **:** Link web.

- Output **:** 1 file raw csv.

2.2. Mô tả dữ liệu

- Số mẫu **:**

- SmallDS : 1000 (samples).
- BigDS : 10000 (samples).

- Chiều dữ liệu **:** 10000 x 16 (gồm 15 đặc trưng và một cột số thứ tự).

- SmallDS : 1000 x 16 (gồm 15 đặc trưng và một cột số thứ tự).
- BigDS : 10000 x 16 (gồm 15 đặc trưng và một cột số thứ tự).

- Số đặc trưng của mẫu**:** 15 đặc trưng.

Bảng 1: Mô tả dữ liệu thô ban đầu của BigDS

Column	Kiểu dữ liệu	Số dữ liệu trống
Rank	Int64	0
Year	Int64	0
Movie	object	0
WorldwideBox Office	object	0
Production Budget	object	5293
Date Releases	object	4710
MPAA	object	2510
Running Time	object	1431
Franchise	object	8036
Genre	object	489
Creative Type	object	739
Production/Financing Companies	object	5329
Production Countries	object	443
Languages	object	1763
Director	object	1297

Bảng 2: Mô tả dữ liệu thô ban đầu của SmallDS

Column	Kiểu dữ liệu	Số dữ liệu trống
Rank	Int64	0
Year	Int64	0
Movie	object	0
WorldwideBox Office	object	0
Production Budget	object	83
Date Releases	object	255
MPAA	object	28
Running Time	object	12
Franchise	object	429
Genre	object	1
Creative Type	object	1
Production/Financing Companies	object	106
Production Countries	object	2
Languages	object	13
Director	object	15

- Dữ liệu SmallDS_raw thu thập được:

```
7 df_small.head()
```

✓ 0.0s Python

	Unnamed: 0	Rank	Year	Movie	WorldwideBox Office	Production Budget	Date Releases	MPAA	Running Time	Franchise	Genre	Creative Type	Production/Financing Companies	Production Countries	Languages	Director
0	0	1	2009	Avatar	$2,923,706,026	$237,000,000	December 17th, 2009	PG-13	162 minutes	Avatar	Action	Science Fiction	Dune Entertainment, 20th Century Fox, Ingeniou...	United States	English, Na'vi	James Cameron
1	1	2	2019	Avengers-Endgame-(2019)	$2,794,731,755	$400,000,000	April 23rd, 2019	PG-13	181 minutes	Marvel Cinematic UniverseAvengers	Action	Super Hero	Marvel Studios	United States	English	Joe Russo,Anthony Russo
2	2	3	2022	Avatar-The-Way-of-Water-(2022)	$2,319,738,066	$460,000,000	December 9th, 2022	PG-13	190 minutes	Avatar	Action	Science Fiction	Lightstorm Entertainment, 20th Century Studios...	United States	English	James Cameron
3	3	4	1997	Titanic-(1997)	$2,222,985,568	$200,000,000	December 18th, 1997	PG-13	194 minutes	NaN	Drama	Historical Fiction	20th Century Fox, Paramount Pictures, Lightsto...	United States	English, Italian, Swedish	James Cameron
4	4	5	2015	Star-Wars-Ep-VII-The-Force-Awakens	$2,064,615,817	$306,000,000	December 16th, 2015	PG-13	136 minutes	Star Wars	Adventure	Science Fiction	Lucasfilm, Bad Robot	United States	English	J.J. Abrams

Hình 4. Dữ liệu BigDS_Raw thu được

- Dữ liệu SmallDS_clean thu được sau khi clean dữ liệu

	Unnamed: 0	Rank	Year	Movie	WorldwideBox Office	Production Budget	Date Releases	MPAA	Running Time	Franchise	Genre	Creative Type	Production/Financing Companies	Pro C
0	0	1	2009	Avatar	2923706026	237000000.0	December 17th, 2009	PG-13	162.0	Avatar	Action	Science Fiction	Dune Entertainment, 20th Century Fox, Ingeniou...	
1	1	2	2019	Avengers-Endgame-(2019)	2794731755	400000000.0	April 23rd, 2019	PG-13	181.0	Marvel Cinematic UniverseAvengers	Action	Super Hero	Marvel Studios	
2	2	3	2022	Avatar-The-Way-of-Water-(2022)	2319738066	460000000.0	December 9th, 2022	PG-13	190.0	Avatar	Action	Science Fiction	Lightstorm Entertainment, 20th Century Studios...	
3	3	4	1997	Titanic-(1997)	2222985568	200000000.0	December 18th, 1997	PG-13	194.0	NaN	Drama	Historical Fiction	20th Century Fox, Paramount Pictures, Lightsto...	
4	4	5	2015	Star-Wars-Ep-VII-The-Force-Awakens	2064615817	306000000.0	December 16th, 2015	PG-13	136.0	Star Wars	Adventure	Science Fiction	Lucasfilm, Bad Robot	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
995	995	996	1995	Dangerous-Minds	178919401	23000000.0	NaN	R	99.0	NaN	Drama	Dramatization	NaN	
996	996	997	2006	Scary-Movie-4-(2006)	178710620	40000000.0	NaN	PG-13	83.0	Scary Movie	Comedy	Contemporary Fiction	NaN	
997	997	998	2003	How-to-Lose-a-Guy-in-10-	178503788	50000000.0	April 18th, 2003	PG-13	116.0	NaN	Romantic Comedy	Contemporary Fiction	NaN	

Hình 5. Dữ liệu SmallDS_clean thu được sau khi clean dữ liệu

2.3. Dán nhãn dữ liệu

- Doanh thu phim sẽ được dán nhãn với các mức doanh thu phim tương ứng.

Mức doanh thu (triệu đô)	Nhãn tương ứng
0-300	1
300-600	2
600-900	3
900-1200	4
1200-1500	5
1500-1800	6
1800-2100	7
2100-2400	8
2400-2700	9
2700-3000	10

Bảng 3: Bảng đánh giá mức doanh thu phim

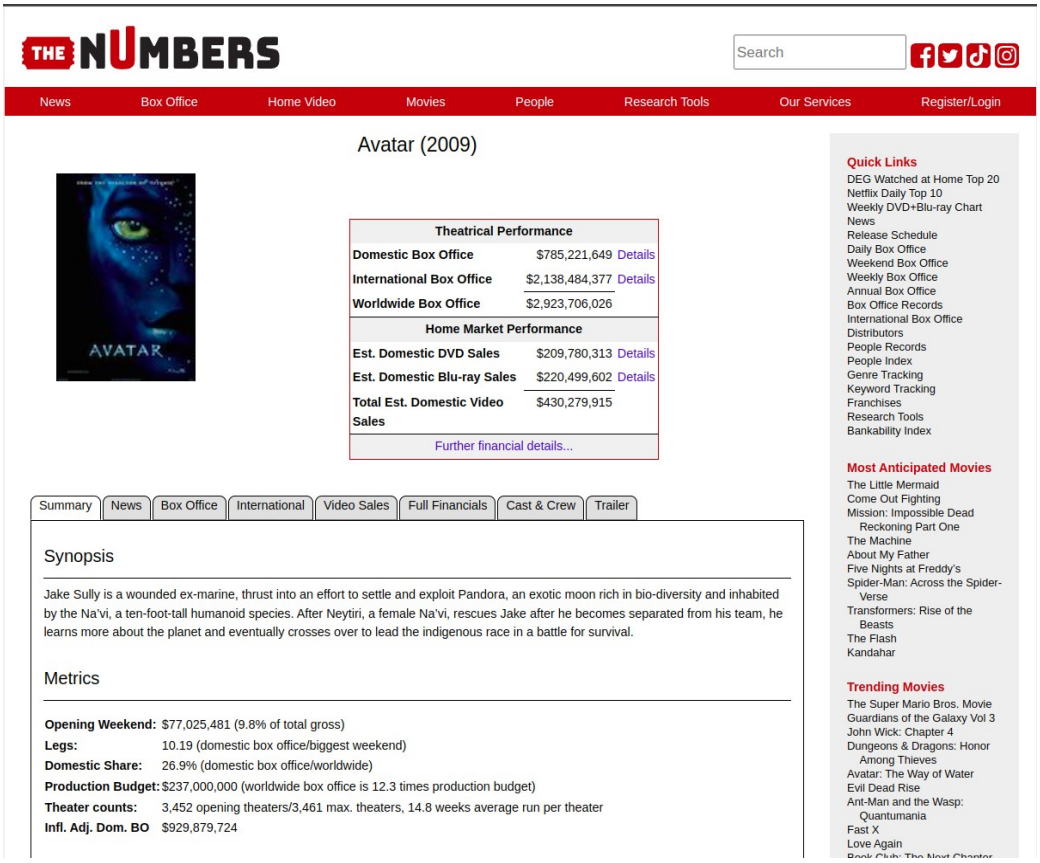
	WorldwideBox Office
105	781947691
68	894860230
479	322459006
399	362605033
434	348901032
...	...
835	209221328
192	560483719
629	258751370
559	288510892
684	245179530

	WorldwideBox Office
105	3
68	3
479	2
399	2
434	2
...	...
835	1
192	2
629	1
559	1
684	1

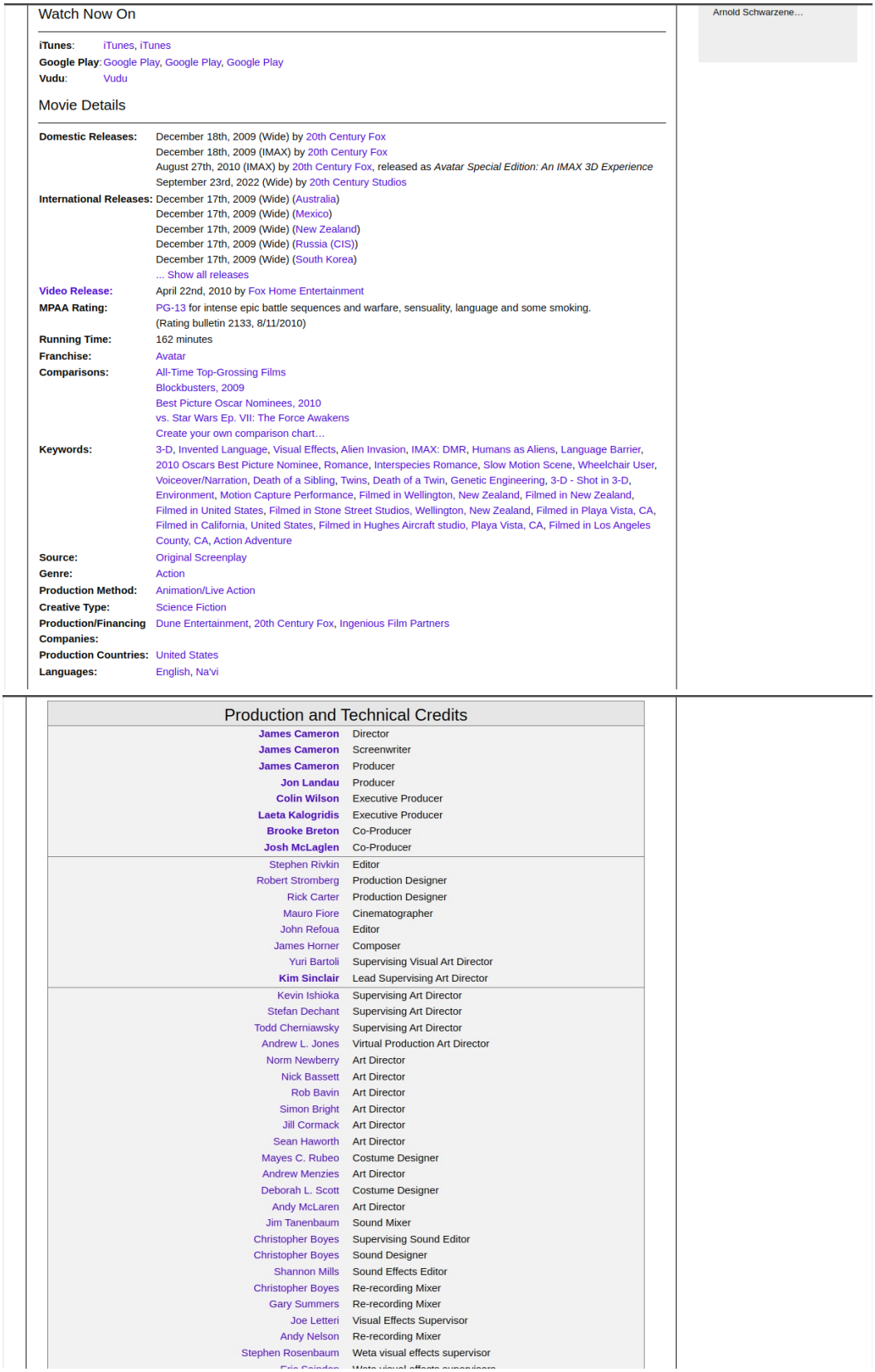
Hình 6. Dữ liệu doanh thu phim trước và sau khi dán nhãn dữ liệu

3. Trích xuất đặc trưng

3.1. Lựa chọn và tạo mới đặc trưng cần thiết



Hình 7. Trích xuất các đặc trưng lựa chọn trong web



Watch Now On

iTunes: iTunes, iTunes
Google Play: Google Play, Google Play, Google Play
Vudu: Vudu

Movie Details

Domestic Releases:	December 18th, 2009 (Wide) by 20th Century Fox December 18th, 2009 (IMAX) by 20th Century Fox August 27th, 2010 (IMAX) by 20th Century Fox, released as *Avatar Special Edition: An IMAX 3D Experience* September 23rd, 2022 (Wide) by 20th Century Studios
International Releases:	December 17th, 2009 (Wide) (Australia) December 17th, 2009 (Wide) (Mexico) December 17th, 2009 (Wide) (New Zealand) December 17th, 2009 (Wide) (Russia (CIS)) December 17th, 2009 (Wide) (South Korea) ... Show all releases
Video Release:	April 22nd, 2010 by Fox Home Entertainment
MPAA Rating:	PG-13 for intense epic battle sequences and warfare, sensuality, language and some smoking. (Rating bulletin 2133, 8/11/2010)
Running Time:	162 minutes
Franchise:	Avatar
Comparisons:	All-Time Top-Grossing Films Blockbusters, 2009 Best Picture Oscar Nominees, 2010 vs. Star Wars Ep. VII: The Force Awakens Create your own comparison chart…
Keywords:	3-D, Invented Language, Visual Effects, Alien Invasion, IMAX: DMR, Humans as Aliens, Language Barrier, 2010 Oscars Best Picture Nominee, Romance, Interspecies Romance, Slow Motion Scene, Wheelchair User, Voiceover/Narration, Death of a Sibling, Twins, Death of a Twin, Genetic Engineering, 3-D - Shot in 3-D, Environment, Motion Capture Performance, Filmed in Wellington, New Zealand, Filmed in New Zealand, Filmed in United States, Filmed in Stone Street Studios, Wellington, New Zealand, Filmed in Playa Vista, CA, Filmed in California, United States, Filmed in Hughes Aircraft studio, Playa Vista, CA, Filmed in Los Angeles County, CA, Action Adventure
Source:	Original Screenplay
Genre:	Action
Production Method:	Animation/Live Action
Creative Type:	Science Fiction
Production/Financing Companies:	Dune Entertainment, 20th Century Fox, Ingenious Film Partners
Production Countries:	United States
Languages:	English, Na'vi

Production and Technical Credits

James Cameron	Director
James Cameron	Screenwriter
James Cameron	Producer
Jon Landau	Producer
Colin Wilson	Executive Producer
Laeta Kalogridis	Executive Producer
Brooke Breton	Co-Producer
Josh McLaglen	Co-Producer
Stephen Rivkin	Editor
Robert Stromberg	Production Designer
Rick Carter	Production Designer
Mauro Fiore	Cinematographer
John Refoua	Editor
James Horner	Composer
Yuri Bartoli	Supervising Visual Art Director
Kim Sinclair	Lead Supervising Art Director
Kevin Ishioka	Supervising Art Director
Stefan Dechant	Supervising Art Director
Todd Cherniawsky	Supervising Art Director
Andrew L. Jones	Virtual Production Art Director
Norm Newberry	Art Director
Nick Bassett	Art Director
Rob Bavin	Art Director
Simon Bright	Art Director
Jill Cormack	Art Director
Sean Haworth	Art Director
Mayes C. Rubeo	Costume Designer
Andrew Menzies	Art Director
Deborah L. Scott	Costume Designer
Andy McLaren	Art Director
Jim Tanenbaum	Sound Mixer
Christopher Boyes	Supervising Sound Editor
Christopher Boyes	Sound Designer
Shannon Mills	Sound Effects Editor
Christopher Boyes	Re-recording Mixer
Gary Summers	Re-recording Mixer
Joe Letteri	Visual Effects Supervisor
Andy Nelson	Re-recording Mixer
Stephen Rosenbaum	Weta visual effects supervisor

Hình 8. Trích xuất các đặc trưng lựa chọn trong web

Các đặc trưng được lựa chọn: Rank, Year, Movie, WorldwideBox Office, Production Budget, Date Releases, MPAA, Running Time, Franchise, Genre, Creative Type, Production/Financing Companies, Production Countries, Languages, Director

Các đặc trưng không được lựa chọn: Reviews, Score, Ratting, Population. Vì mục tiêu của nhóm là dự đoán mức doanh thu khi phim còn trong giai đoạn ý tưởng, hay vừa mới ra mắt, các đặc trưng này chưa tồn tại ở thời điểm đó.

Các đặc trưng được tạo thêm: Từ Date Release tạo ra hai đặc trưng mới là Day Release và Month Release.

3.2 Làm sạch dữ liệu với đúng kiểu dữ liệu

- Sau khi đã lựa chọn các đặc trưng thì cào dữ liệu theo các thẻ HTML tương ứng.

- Loại bỏ các giá trị NaN bằng hàm mode().

- Loại bỏ các kí tự đặc biệt và chữ trong cột: WorldwideBox Office, Production Budget, Date Releases, Running Time.

- Các đặc trưng WorldwideBox Office, Production Budget sau khi xử lý và từ đặc trưng Date Releases tạo thêm hai đặc trưng mới Day Releases và Month Release

nk	Year	Movie	WorldwideBox Office	Production Budget	Date Releases	MPAA	Running Time	r	Day Releases	Month Releases
1	2009	Avatar	2923706026	237000000.0	December 17th, 2009	PG-13	162.0	s	3.0	12.0
2	2019	Avengers-Endgame-(2019)	2794731755	400000000.0	April 23rd, 2019	PG-13	181.0	Ne Uy	1.0	4.0
3	2022	Avatar-The-Way-of-Water-(2022)	2319738066	460000000.0	December 9th, 2022	PG-13	190.0	s	4.0	12.0
4	1997	Titanic-(1997)	2222985568	200000000.0	December 18th, 1997	PG-13	194.0	s	3.0	12.0
5	2015	Star-Wars-Ep-VII-The-Force-Awakens	2064615817	306000000.0	December 16th, 2015	PG-13	136.0	s	2.0	12.0

Hình 9. Các đặc trưng WorldwideBox Office, Production Budget, Date Release sau khi xử lý

- Các đặc trưng Rank, WorldwideBox Office, Production Budget, Running Time sau khi xử lý kiểu dữ liệu

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 18 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Unnamed: 0                      1000 non-null   int64
 1   Rank                            1000 non-null   object
 2   Year                            1000 non-null   int64
 3   Movie                           1000 non-null   object
 4   WorldwideBox Office             1000 non-null   int64
 5   Production Budget               917 non-null    float64
 6   Date Releases                   745 non-null    object
 7   MPAA                            972 non-null    object
 8   Running Time                    988 non-null    float64
 9   Franchise                       571 non-null    object
 10  Genre                           999 non-null    object
 11  Creative Type                   999 non-null    object
 12  Production/Financing Companies  894 non-null    object
 13  Production Countries            998 non-null    object
 14  Languages                       987 non-null    object
 15  Director                        985 non-null    object
 16  Day Releases                    745 non-null    float64
 17  Month Releases                  745 non-null    float64
dtypes: float64(4), int64(3), object(11)
memory usage: 140.8+ KB
```

Hình 10. Các đặc trưng WorldwideBox Office, Production Budget sau khi xử lý kiểu dữ liệu

3.3 Mã hóa dữ liệu

- Các bộ phim có đạo diễn thuộc top 100 đạo diễn có doanh thu cao nhất sẽ được gán là 1 ngược lại là 0.

- Các bộ phim có công ty sản xuất thuộc top 100 công ty có doanh tu cao nhất sẽ được gán là 1 ngược lại là 0.

- Các bộ phim thuộc top 15 series có doanh thu cao nhất sẽ được gán là 1 ngược lại là 0, nếu không thuộc series nào sẽ được gán là 2.

- Các đặc trưng kiểu category khác được mã hóa bằng LabelEncoder.

Rank	Year	Movie	WorldwideBox Office	Production Budget	Date Releases	MPAA	Running Time	Franchise	Genre	Creative Type	Production/Financing Companies	Production Countries	Languages	Director	Day Releases	Month Releases
106	2016	Deadpool	781947691	58000000.0	February 9th, 2016	R	107.0	X-MenDeadpool	Action	Super Hero	[Marvel Studios]	[United States]	[English]	[Tim Miller]	1.0	2.0
69	2007	Spider-Man-3	894860230	258000000.0	NaN	PG-13	139.0	Spider-Man	Adventure	Super Hero	[Columbia Pictures, Marvel Studios, Laura Zisk...	[United States]	[English]	[Sam Raimi]	NaN	NaN
480	2010	Robin-Hood-(2010)	322459006	210000000.0	NaN	PG-13	139.0	NaN	Action	Historical Fiction	[United Artists, Fairbanks]	[United Kingdom, United States]	[English]	[Ridley Scott]	NaN	NaN
400	2007	Alvin-and-the-Chipmunks-(2007)	362605033	55000000.0	NaN	PG	91.0	Alvin and the Chipmunks	Adventure	Kids Fiction	[Regency Enterprises, Fox 2000 Pictures, Bagda...	[United States]	[English]	[Tim Hill]	NaN	NaN
435	2018	Mary-Poppins-Returns-(2018)	348901032	130000000.0	December 20th, 2018	PG	130.0	Mary Poppins	Musical	Kids Fiction	[Walt Disney Pictures, Lucamar, Marc Platt Pro...	[United States]	[English]	[Rob Marshall]	3.0	12.0

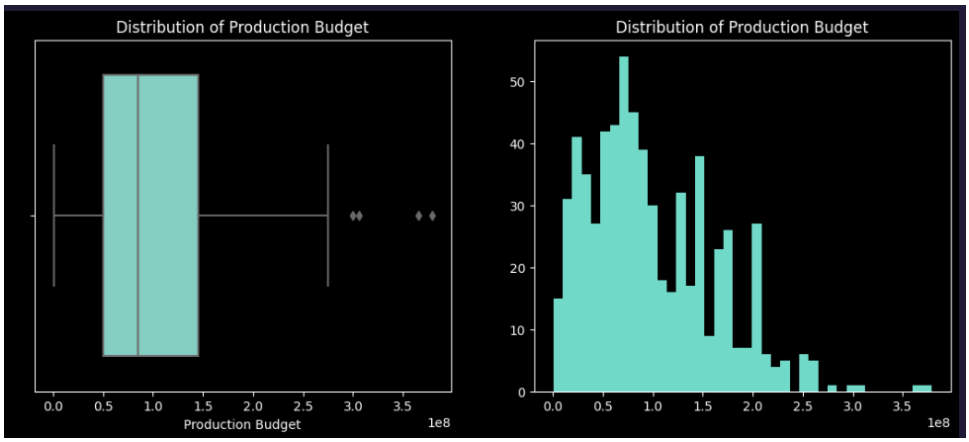
Hình 11. Dữ liệu huấn luyện trước khi được mã hóa

	Production Budget	MPAA	Running Time	Genre	Creative Type	topCompanies	topDirector	Day Releases	Month Releases	holiday_month	origin_isEnglish	topFranchise	countries_isUS
105	58000000.0	7	107.0	0	8	1	1	1.0	2.0	0	1	2	1
68	258000000.0	4	139.0	1	8	1	1	4.0	6.0	0	1	2	1
479	210000000.0	4	139.0	0	4	1	1	4.0	6.0	0	1	0	1
399	55000000.0	3	91.0	1	5	1	0	4.0	6.0	0	1	2	1
434	130000000.0	3	130.0	7	5	1	1	3.0	12.0	1	1	2	1

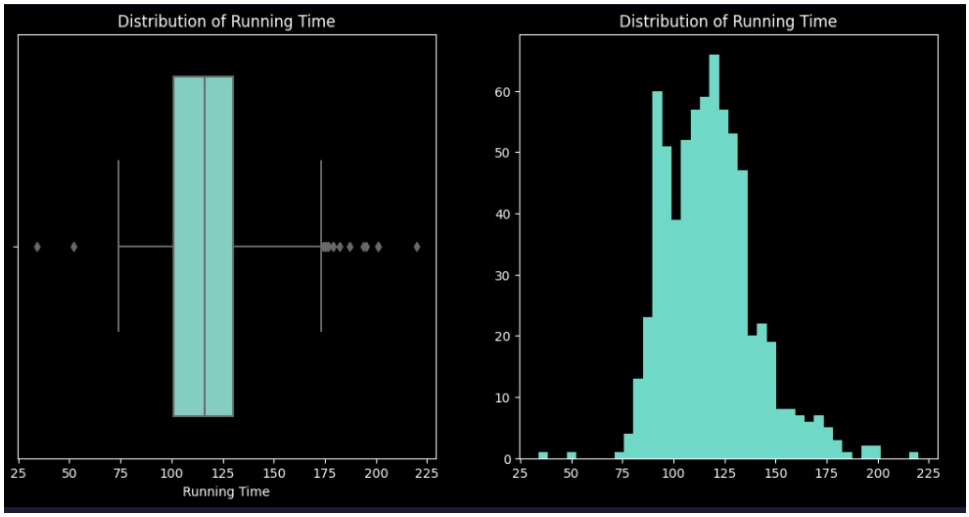
Hình 12. Dữ liệu huấn luyện sau khi được mã hóa

3.4 Xử lý ngoại lệ

- Sau khi chia bộ dữ liệu thành 2 tập huấn luyện và kiểm thử với kích thước tập kiểm thử là 30%. Ta vẽ biểu đồ boxplot để kiểm tra ngoại lệ của 2 đặc trưng trong tập dữ liệu huấn luyện.



Hình 13. Phân bố dữ liệu huấn luyện của đặc trưng Production Budget khi chưa xử lý ngoại lệ



Hình 14. Phân bố dữ liệu huấn luyện của đặc trưng Running Time khi chưa xử lý ngoại lệ

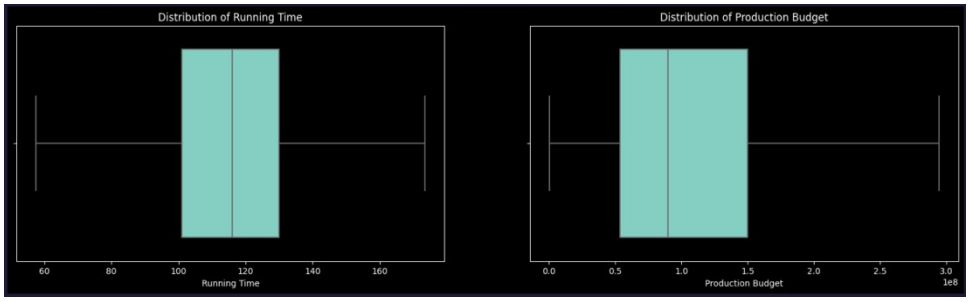
- Hai đặc trưng trên đều có phân bố lệch, nên ta sẽ sử dụng IQR để tìm biên cho phần xử lý ngoại lệ.

- Xử lí ngoại lệ với phân bố lệch:

+ Biên trên = quantile(0.75) + 1.5 * iqr

+ Biên dưới = quantile(0.25) - 1.5* iqr

- Kết quả phân bố dữ liệu sau khi xử lý ngoại lệ được thể hiện qua biểu đồ sau.



Hình 15. Phân bố dữ liệu huấn luyện sau khi xử lý ngoại lệ

3.5 Chuẩn hóa dữ liệu

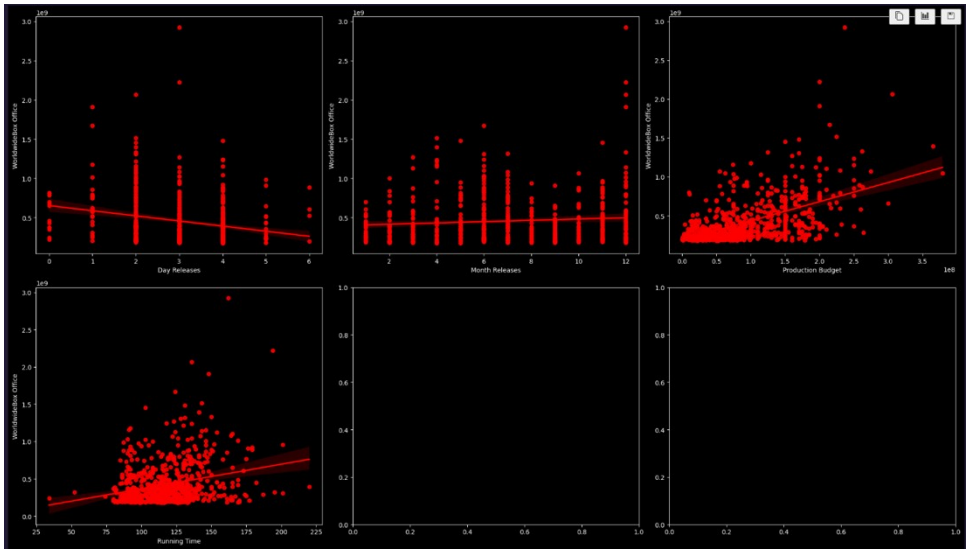
- Chuẩn hóa dữ liệu sử dụng StandardScaler của thư viện sklearn với thuộc tính fit_transform cho tập dữ liệu huấn luyện và tập dữ liệu kiểm thử.

	Production Budget	MPAA	Running Time	Genre	Creative Type	topCompanies	topDirector	Day Releases	Month Releases	holiday_month	origin_isEnglish	topFranchise	countries_isUS
105	58000000.0	7	107.0	0	8	1	1	1.0	2.0	0	1	2	1
68	258000000.0	4	139.0	1	8	1	1	4.0	6.0	1	1	2	1
479	210000000.0	4	139.0	0	4	1	1	4.0	6.0	1	1	0	1
399	55000000.0	3	91.0	1	5	1	0	4.0	6.0	1	1	2	1
434	130000000.0	3	130.0	7	5	1	1	3.0	12.0	1	1	2	1

Hình 16. Dữ liệu huấn luyện sau khi chuẩn hóa dữ liệu

3.6 Độ tương quan giữa các đặc trưng với biến mục tiêu

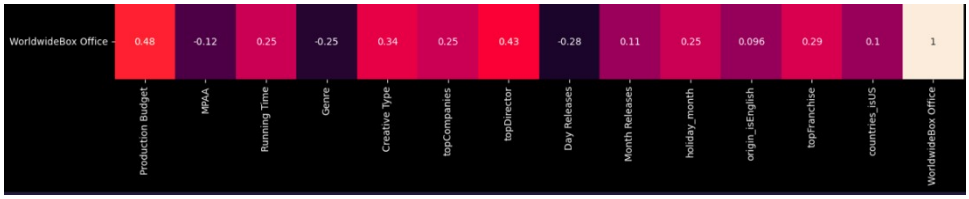
- Tiến hành trực quan hóa sự tương quan của các đặc trưng dạng số so với đặc trưng mục tiêu là ‘WorldwideBox Office’ bằng sơ đồ regplot. Ta thu được sơ đồ dưới đây:



Hình 17. Sự tương quan giữa các đặc trưng với biến mục tiêu

- Có thể đặc trưng ‘Production Budget’, ‘Running Time’ có độ tương quan khá cao. Khi giá trị của các đặc trưng này tăng thì doanh thu phim cũng tăng. Ngoài ra, đặc trưng ‘Day

Realeases’ cũng có sự tương quan nhẹ. Còn đặc trưng ‘Month Releases’ có độ tương quan không cao.



Hình 18. Sự tương quan giữa các đặc trưng với biến mục tiêu

```
WorldwideBox Office    1.000000
Production Budget      0.484951
topDirector            0.430321
Creative Type          0.343442
topFranchise           0.291902
topCompanies           0.250068
Running Time           0.249773
holiday_month          0.246166
Month Releases         0.111579
countries_isUS         0.104988
origin_isEnglish       0.096471
MPAA                  -0.118873
Genre                 -0.252266
Day Releases          -0.280336
Name: WorldwideBox Office, dtype: float64
```

Hình 19. Độ tương quan giảm dần của các đặc trưng so với biến mục tiêu

- Có thể thấy ‘Production Budget’ có độ tương quan với doanh thu cao nhất. Bộ dữ liệu có các đặc trưng với mức độ tương quan khá thấp cao so với đặc trưng mục tiêu là ‘WorldwideBox Office’- Doanh thu phim

- Tổng cộng có 12 đặc trưng được lựa chọn cho mô hình dự đoán: ‘Production Budget’, ‘MPAA’, ‘Running Time’, ‘Genre’, ‘Creative Type’, ‘topCompanies, ‘Top Director’, ‘Day Release’, ‘Month Release’, ‘holiday_month’, ‘origin_isEnglish’, ‘topFranchise’, ‘countries_isUS’.

4. Mô hình hóa dữ liệu

4.1 Các metrics dùng để đánh giá:

a) Độ chính xác (accuracy):

- Tỷ lệ các trường hợp dự đoán đúng trên tổng số các trường hợp.

$$Accuracy = \frac{TP + TN}{total\ sample}$$

b) Precision:

- Tỷ lệ dự đoán positive đúng trên tổng số trường hợp positive dự đoán.

$$Pr\,esicion = \frac{TP}{total\ predicted\ positive} = \frac{TP}{TP + FP}$$

c) Recall:

- Tỷ lệ dự đoán positve đúng trên tổng số trường hợp positive thực.

$$Recall = \frac{TP}{total\ actual\ positive} = \frac{TP}{TP + FN}$$

d) F1 Score:

- Trung bình điều hòa giữa precision và recall. Do đó nó đại diện hơn trong việc đánh giá độ chính xác trên đồng thời precision và recall.

$$F_1 = \frac{2}{precisio\,n^{-1} + recal\,l^{-1}}$$

e) Các chỉ số TP, FP, TN, FN lần lượt có ý nghĩa là:

- TP (True Positive): Tổng số trường hợp dự báo khớp Positive.
- TN (True Negative): Tổng số trường hợp dự báo khớp Negative.
- FP (False Positive): Tổng số trường hợp dự báo các quan sát thuộc nhãn Negative thành Positive.
- FN (False Negative): Tổng số trường hợp dự báo các quan sát thuộc nhãn Positive thành Negative.

Positive và Negative được qui ước là 1 và 0

4.2 Mô hình Logictic Regression:

a) Cơ sở lý thuyết:

Hồi quy logistic là một kỹ thuật data analysis thường là toán học để xác định sự liên quan giữa hai yếu tố của data. Sau đó, dựa vào sự liên quan này để tiến hành đưa ra các dự đoán của 1 yếu tố khi cho biết yếu tố còn lại.

Giống như tất cả các phân tích Regression, Regression logistic là một phân tích dự đoán. Regression logistic được sử dụng để mô tả dữ liệu và giải thích mối quan hệ giữa một biến nhị phân phụ thuộc và một hoặc nhiều biến độc lập cấp danh nghĩa, thứ tự, khoảng hoặc tỷ lệ.

Hồi quy logistic là một kỹ thuật quan trọng trong lĩnh vực trí tuệ nhân tạo và máy học (AI/ML). Mô hình ML là các chương trình phần mềm có thể được đào tạo để thực hiện các tác vụ xử lý dữ liệu phức tạp mà không cần sự can thiệp của con người. Mô hình ML được xây dựng bằng hồi quy logistic có thể giúp các tổ chức thu được thông tin chuyên sâu hữu ích từ dữ liệu kinh doanh của mình. Họ có thể sử dụng những thông tin chuyên sâu này để phân tích dự đoán nhằm giảm chi phí hoạt động, tăng độ hiệu quả và đổi chỉnh quy mô nhanh hơn.

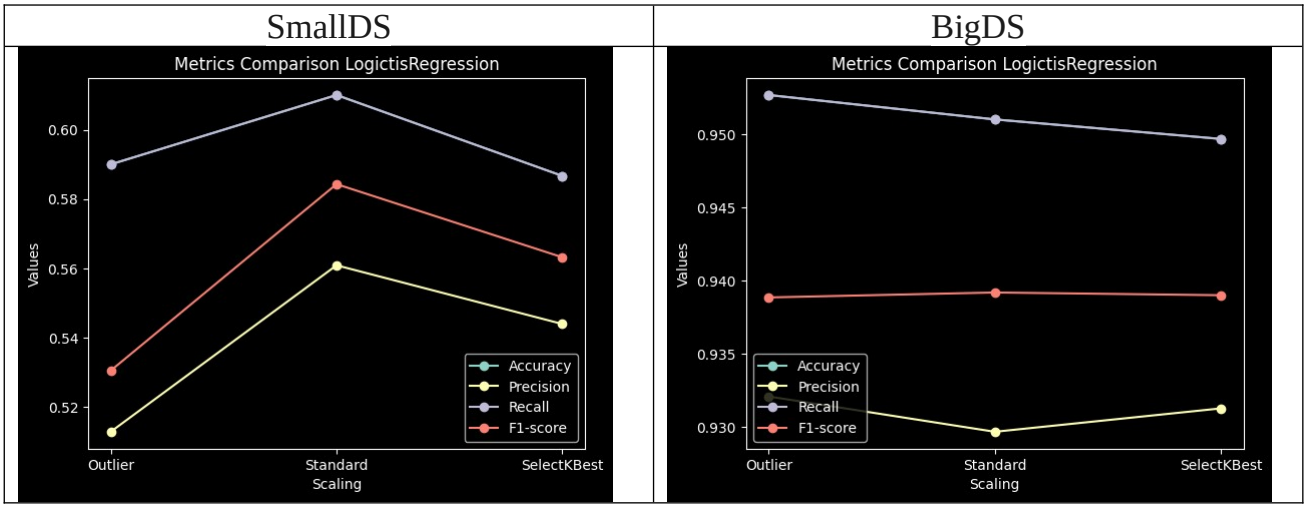
b) Chia dữ liệu và train mô hình

- Bộ tham số của mô hình: Sử dụng thuật toán RandomizedSearchCV để tìm bộ siêu tham số tối ưu cho mô hình.

Tên tham số	Mảng giá trị	Ý nghĩa
C	uniform(loc=0, scale=4)	Tham số này quy định độ mạnh mẽ của regularization. Càng nhỏ, mô hình càng được regularization mạnh hơn và ngược lại.
penalty	['l1', 'l2']	Tham số này quy định phương pháp sử dụng để xử lý đa cộng tuyến
solver	['liblinear', 'saga']	Tham số này xác định phương pháp sử dụng để tìm nghiệm của mô hình

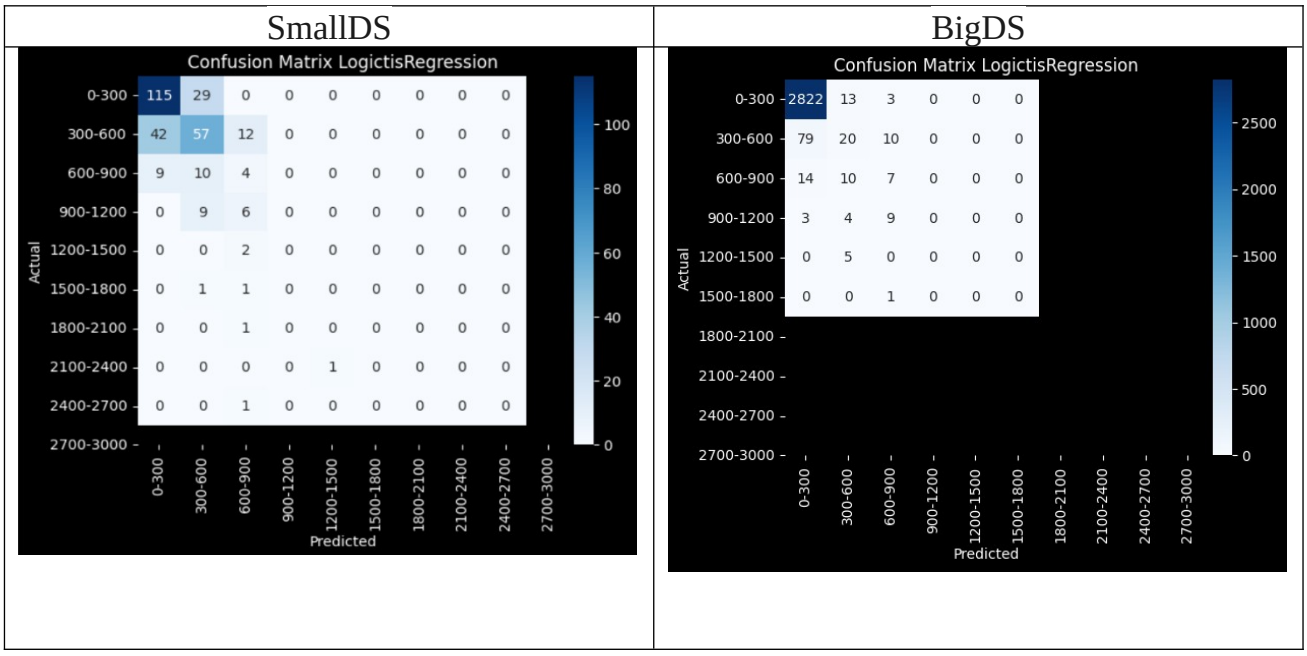
- Chia tập dữ liệu Huấn luyện/Kiểm thử: theo mô hình 30% kiểm thử.

c) Độ thị thể hiện hiệu suất



Hình 20. Đồ thị thể hiện hiệu suất 4 metrics trên mô hình logistic regresstion

d) Đồ thị dự đoán so với thực tế



Hình 21. Ma trận nhầm lẫn kết quả dự đoán với kết quả thực của test của mô hình Logistic Regresstion

e) Đánh giá mô hình theo các metrics là Accuary, Precesion, Recall, F1-Score

SmallDS

	Accuracy	Precision	Recall	F1-score
Outlier	0.593333	0.513710	0.593333	0.532991
Standard	0.606667	0.581959	0.606667	0.592468
SelectKBest	0.613333	0.551845	0.613333	0.577135

BigDS

	Accuracy	Precision	Recall	F1-score
Outlier	0.952667	0.932077	0.952667	0.938840
Standard	0.951000	0.929664	0.951000	0.939178
SelectKBest	0.949667	0.931263	0.949667	0.938991

Hình 22. Đánh giá mô hình Logistic Regresstion theo các metrics là Accuary, Precesion, Recall, F1-Scall

4.3 Mô hình Random Forest

a) Cơ sở lý thuyết

Random forest là thuật toán supervised learning, có thể giải quyết cả bài toán regression và classification

Random là ngẫu nhiên, Forest là rừng, nên ở thuật toán Random Forest mình sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.

b) Chia dữ liệu và train mô hình

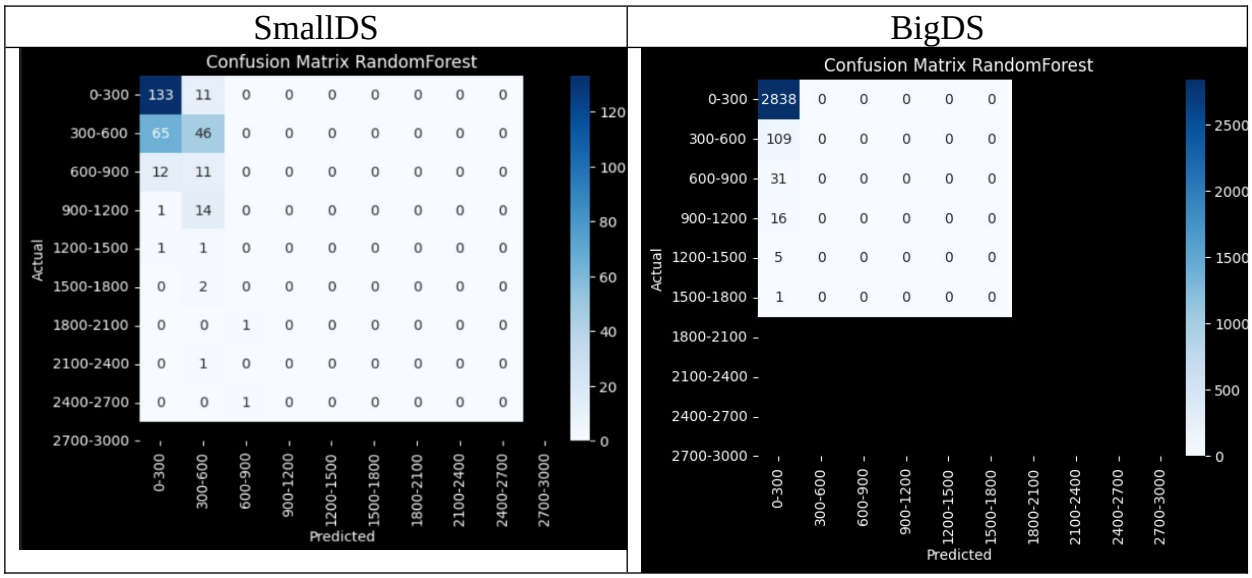
- Bộ tham số của mô hình: Sử dụng thuật toán RandomizedSearchCV để tìm bộ siêu tham số tối ưu cho mô hình.

Tên tham số	Mảng giá trị	Ý nghĩa
n_estimators	[10, 50, 100]	Số lượng cây trong rừng (forest)
max_depth	[None, 5, 10]	Độ sâu tối đa của các cây trong rừng. Nếu không được thiết lập, các cây sẽ được phát triển cho đến khi tất cả các lá đều thuần nhất hoặc số mẫu tối thiểu cần thiết để chia tiếp tục không thể đạt được.
min_samples_split	uniform(loc=0, scale=1)	Số lượng mẫu tối thiểu cần thiết để chia một nút trong cây. Nếu một nút

		có số lượng mẫu ít hơn giá trị này, nó sẽ không được chia
min_samples_leaf	uniform(loc=0, scale=0.5)	Số lượng mẫu tối thiểu cần thiết để tạo một lá trong cây. Nếu một lá có số lượng mẫu ít hơn giá trị này, nó sẽ không được tạo

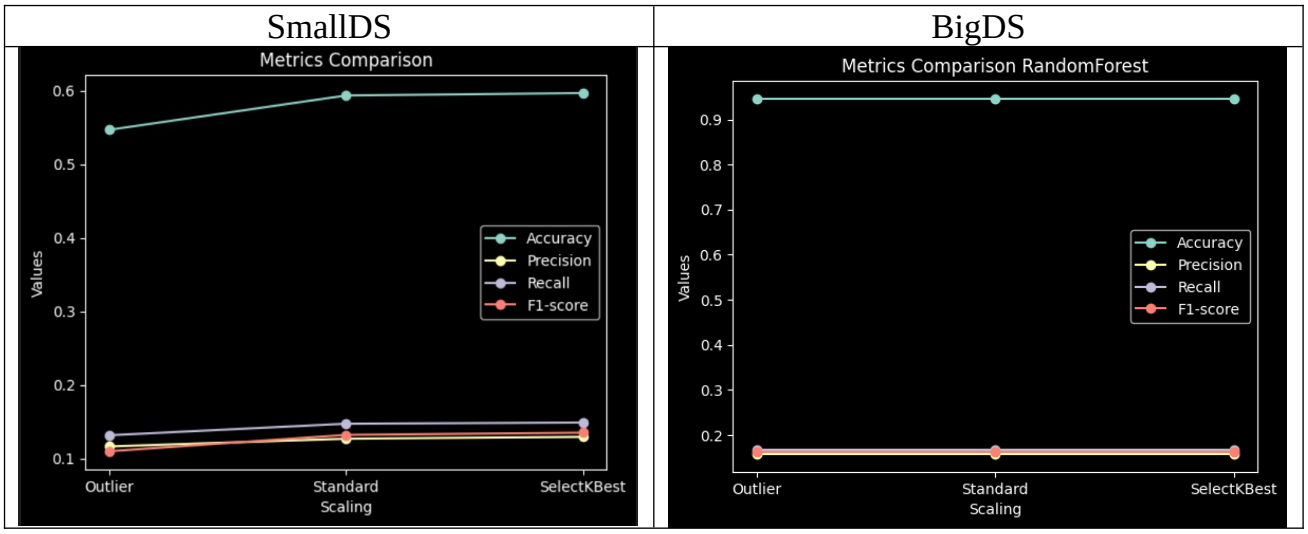
- Chia tập dữ liệu Huấn luyện/Kiểm thử: theo mô hình 30% kiểm thử.

c) Độ lệch dự đoán so với thực tế



Hình 23. Ma trận nhầm lẫn kết quả dự đoán với kết quả thực của test của mô hình Random Forest

d) Đồ thị thể hiện hiệu suất



Hình 24. Đồ thị thể hiện hiệu suất của 4 metrics trên mô hình Random Forest

e) Đánh giá mô hình theo các metrics là Accuary, Precesion, Recall, F1-Score

SmallDS

	Accuracy	Precision	Recall	F1-score
Outlier	0.570000	0.119874	0.139285	0.121892
Standard	0.576667	0.123839	0.140828	0.123154
SelectKBest	0.570000	0.119444	0.138597	0.119863

BigDS

	Accuracy	Precision	Recall	F1-score
Outlier	0.946	0.157667	0.166667	0.162042
Standard	0.946	0.157667	0.166667	0.162042
SelectKBest	0.946	0.157667	0.166667	0.162042

Hình 25. Đánh giá mô hình theo các metrics là Accuary, Precesion, Recall, F1-Scall theo Random Forest

5. Kết luận

Những việc đã làm và kết quả:

- Đã cào được dữ liệu từ web.
- Trích xuất và chuẩn hoá dữ liệu.
- Xây dựng và kiểm thử hai mô hình.
- Đánh giá doanh thu dự đoán được từ hai mô hình Logistic Regression và Random Forest:
  - Với SmallDS cả hai mô đều cho kết quả dự đoán trung bình.
  - Với BigDS cả hai mô hình cho ra kết quả dự đoán chính xác hơn rất nhiều.
  - Mô hình Logistic Resgresstion sẽ phù hợp hơn so với Random Forest.

Hướng phát triển:

- Nếu cào thêm đặt trưng Reviews, Score, Ratting, Population thì sẽ tăng mạnh độ chính xác của mô hình, vì các bộ phim điểm cao thì sẽ có chất lượng tốt và được xem nhiều hơn. Tuy nhiên nếu áp dụng đặt trưng này vào mô hình thì sau này x_predict cũng sẽ cần đặt trưng các đặc trưng Reviews, Score, Ratting, Population, đồng nghĩa với việc bộ phim đã phải hoàn thành đánh giá điểm số (đi ngược lại với mục đích ban đầu của nhóm, đó là dự đoán khi bộ phim còn trong giai đoạn lên ý tưởng).

6. Tài liệu tham khảo

[1] Logistic Regression - Bài toán cơ bản trong Machine Learning (viblo.asia)

[2] What is Logistic regression? | IBM

[3] Random Forest algorithm — Machine Learning cho dữ liệu dạng bảng (machinelearningcoban.com)

[4] What is Random Forest? | IBM

[5] Web Crawling in Python - MachineLearningMastery.com