

TRƯỜNG ĐẠI HỌC BÁCH KHOA - ĐẠI HỌC ĐÀ NẴNG
KHOA CÔNG NGHỆ THÔNG TIN



DỰ ĐOÁN MỨC DOANH THU PHIM

Thành viên nhóm:
Nguyễn Đức Quốc
Nguyễn Trần Thảo Vy
Lê Hoàng

Phân chia công việc

MSSV	Họ Và Tên	Công Việc
102200108	Nguyễn Đức Quốc	<ul style="list-style-type: none">• Làm sạch dữ liệu, xử lý dữ liệu trống• Xử lý ngoại lệ• Mã hóa dữ liệu• Lựa chọn đặc trưng
102200121	Nguyễn Trần Thảo Vy	<ul style="list-style-type: none">• Thu thập dữ liệu• Thống kê mô tả trực quan về dữ liệu• Dán nhãn dữ liệu
102200089	Lê Hoàng	<ul style="list-style-type: none">• Mô hình hóa dữ liệu• Tìm bộ siêu tham số cho mô hình• Chuẩn hóa dữ liệu• Đánh giá hiệu quả mô hình



Tổng quan

1

Mục tiêu và giải pháp

3

Trích xuất đặc trưng

2

Thu thập và mô tả dữ liệu

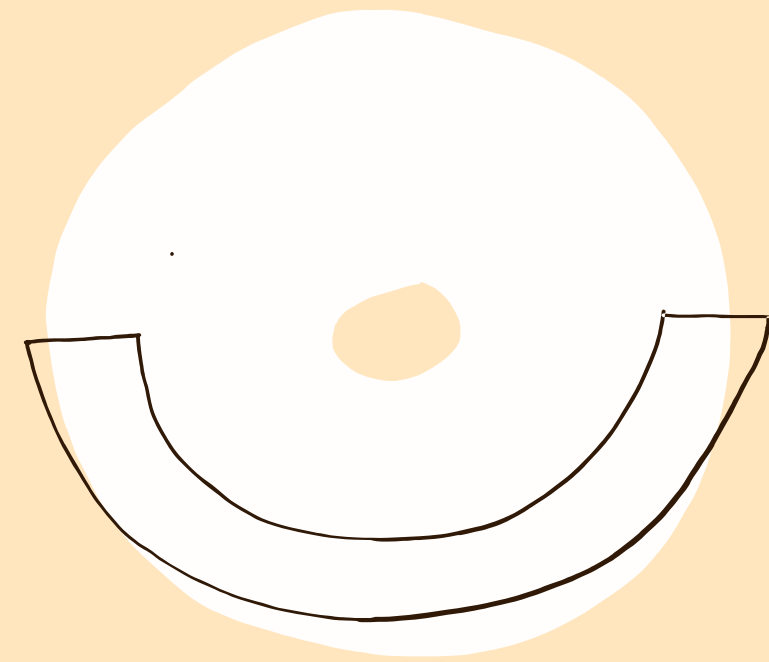
4

Mô hình hóa dữ liệu

5

Kết luận và hướng phát triển

MỤC TIÊU VÀ GIẢI PHÁP



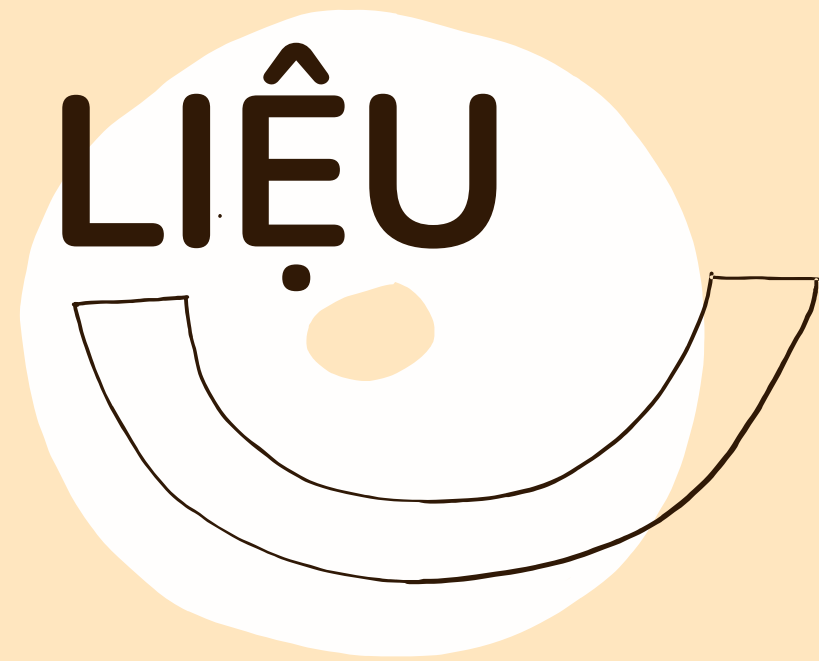
Mục tiêu:

- Dự đoán mức doanh thu phim

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.

THU THẬP VÀ MÔ TẢ DỮ LIỆU



Thu thập dữ liệu:

- Nguồn dữ liệu: <https://www.the-numbers.com/>
- Công cụ thu thập:
 - IDE: Visual Studio Code
 - Ngôn ngữ: Python
 - Thư viện: BeautifulSoup4, request

THU THẬP VÀ MÔ TẢ DỮ LIỆU

Thu thập dữ liệu:

- Nguồn dữ liệu: <https://www.the-numbers.com/>
- Dữ liệu thu thập được:
- Dữ liệu SmallDS: kích thước 1000 x 16 (15 đặc trưng và một cột thứ tự)
- Dữ liệu BigDS: kích thước 10000 x 16 (15 đặc trưng và một cột thứ tự)

7 df_small.head()

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

THU THẬP VÀ MÔ TẢ DỮ LIỆU

Mô tả dữ liệu thô SmallDS:

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
<u>WorldwideBox Office</u>	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

THU THẬP VÀ MÔ TẢ DỮ LIỆU

Dữ liệu SmallDS 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	Pr C
0	0	1	2009	Avatar	2923706026	2370000000.0	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	4000000000.0	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	4600000000.0	December 9th, 2022	PG-13	190.0	Avatar	Action	Science Fiction	Lightstorm Entertainment, 20th Century Studios...
3	3	4	1997	Titanic-(1997)	2222985568	2000000000.0	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	3060000000.0	December 16th, 2015	PG-13	136.0	Star Wars	Adventure	Science Fiction	Lucasfilm, Bad Robot
...
995	995	996	1995	Dangerous-Minds	178919401	230000000.0	NaN	R	99.0	NaN	Drama	Dramatization	NaN
996	996	997	2006	Scary-Movie-4-(2006)	178710620	400000000.0	NaN	PG-13	83.0	Scary Movie	Comedy	Contemporary Fiction	NaN
997	997	998	2003	How-to-Lose-a-Guy-in-10-	178503788	500000000.0	April 18th, 2003	PG-13	116.0	NaN	Romantic Comedy	Contemporary Fiction	NaN

THU THẬP VÀ MÔ TẢ DỮ LIỆU

Dữ liệu doanh thu trước và sau khi được dán nhãn

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 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

TRÍCH XUẤT ĐẶC TRƯNG

Lựa chọn và tạo mới đặc trưng cần thiết

- 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, Rating, 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.

TRÍCH XUẤT ĐẶC TRƯNG

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.

TRÍCH XUẤT ĐẶC TRƯNG

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

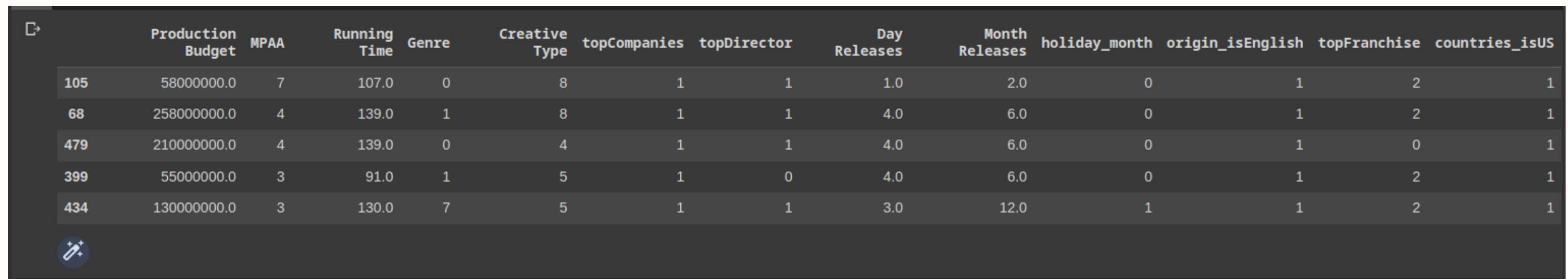
Các đặc trưng WorldwideBox Office, Production Budget, Date Release sau khi xử lý

```
<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
```

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

TRÍCH XUẤT ĐẶC TRƯNG

- 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 thu 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.



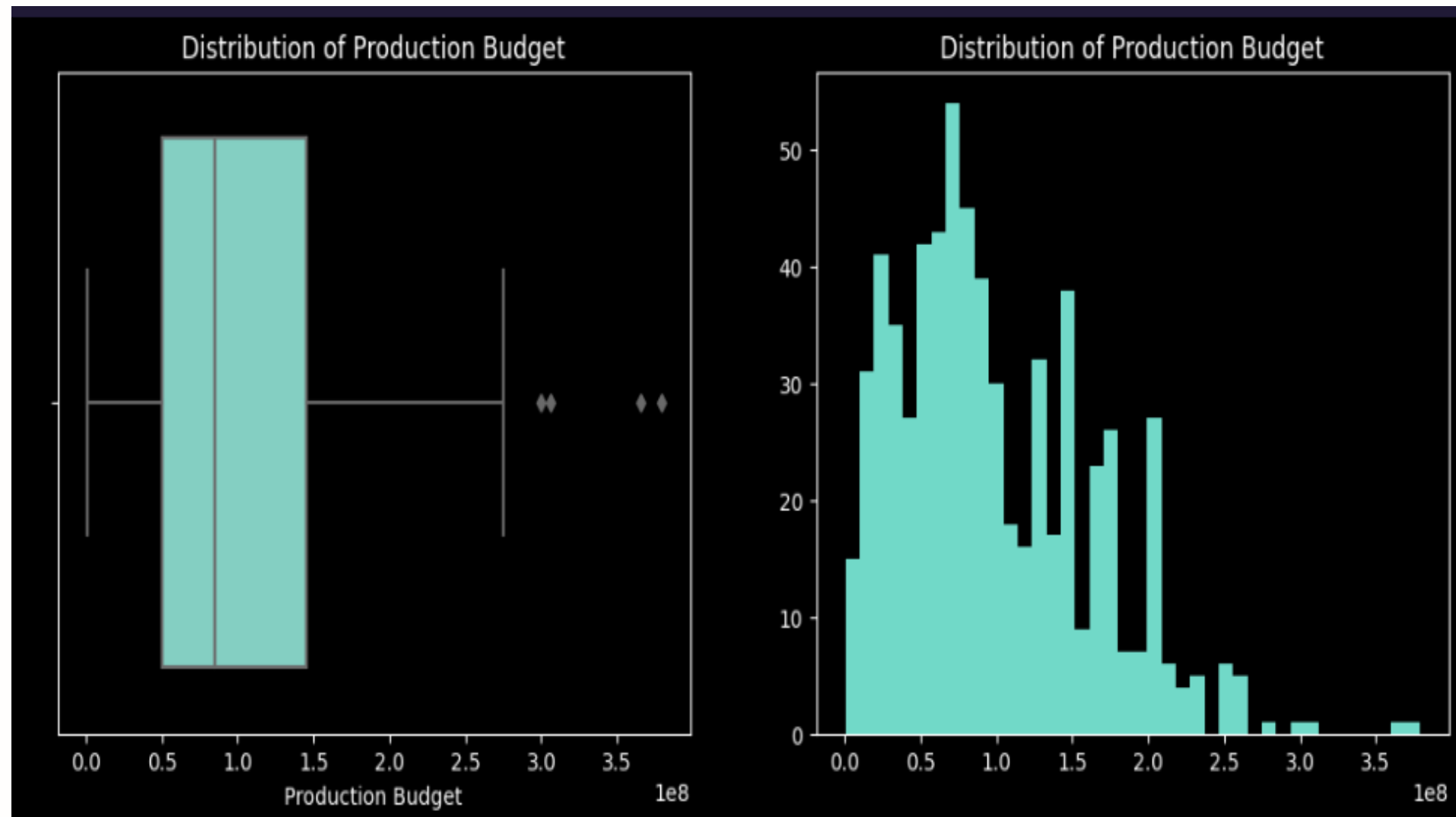
The screenshot shows a Jupyter Notebook interface with a table of movie data. The table has 15 columns: an index column, Production Budget, MPAA, Running Time, Genre, Creative Type, topCompanies, topDirector, Day Releases, Month Releases, holiday_month, origin_isEnglish, topFranchise, and countries_isUS. The data is presented in a dark-themed environment with a light gray header row.

	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

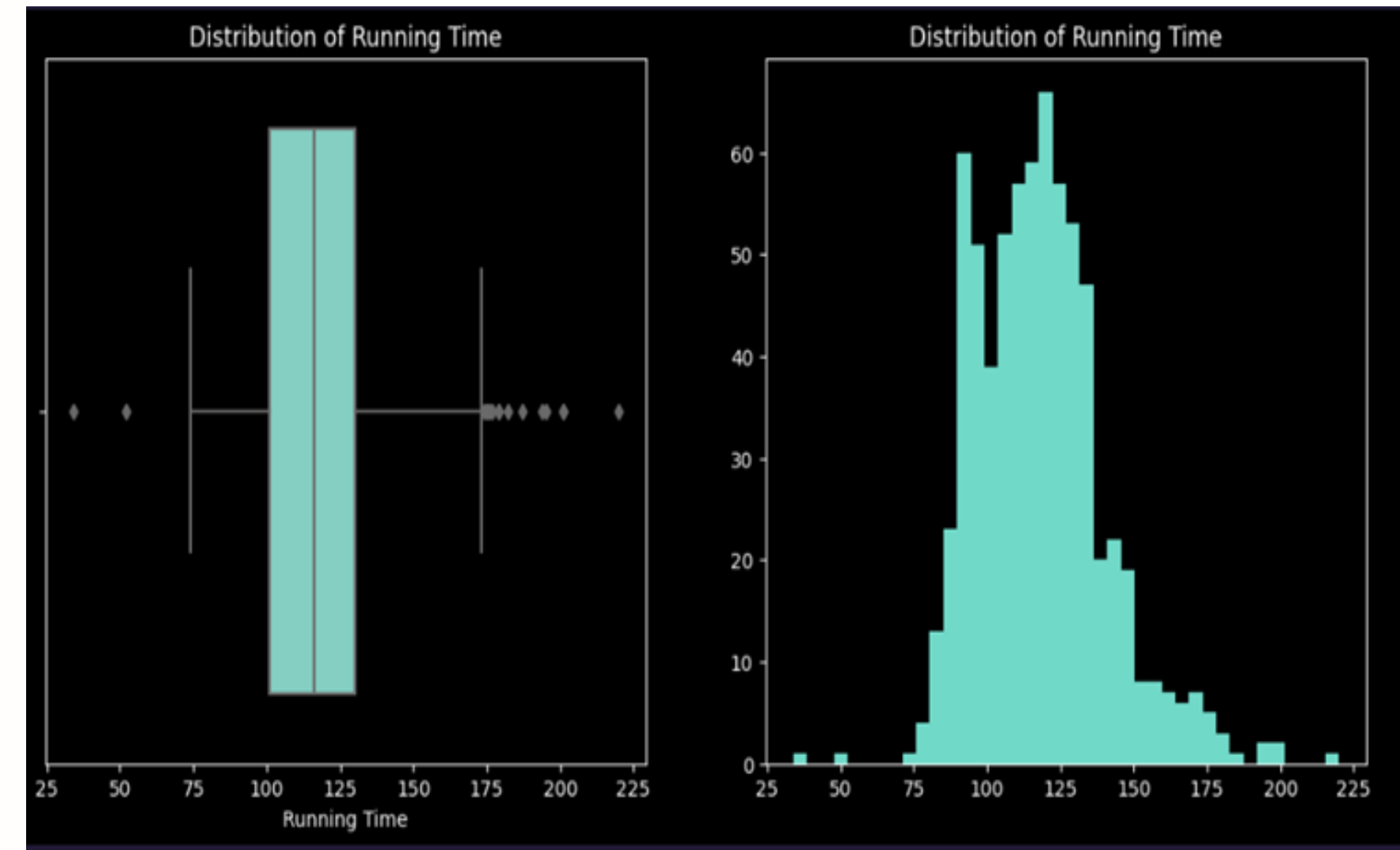
Dữ liệu huấn luyện sau khi được mã hóa

TRÍCH XUẤT ĐẶC TRƯNG

- 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 số trong tập dữ liệu huấn luyện.



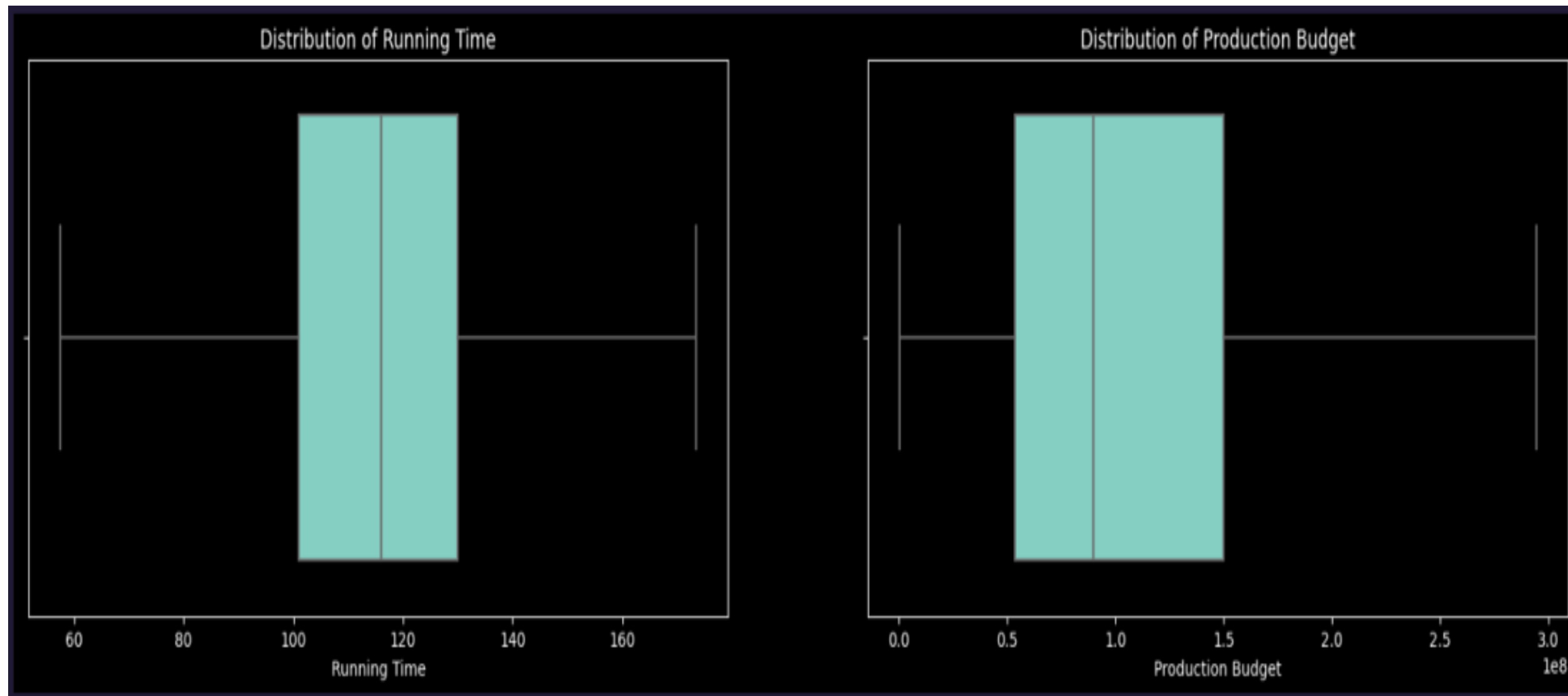
Phân bố dữ liệu huấn luyện của đặc trưng Production Budget khi chưa xử lý ngoại lệ



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

TRÍCH XUẤT ĐẶC TRƯNG

- Hai đặc trưng trên đều có phân bố lệch, nên ta sẽ sử dụng IQR để tìm biên cho phần xử lý ngoại lệ. Kết quả phân bố dữ liệu sau khi xử lý ngoại lệ được thể hiện qua biểu đồ sau.



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

TRÍCH XUẤT ĐẶC TRƯNG

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.

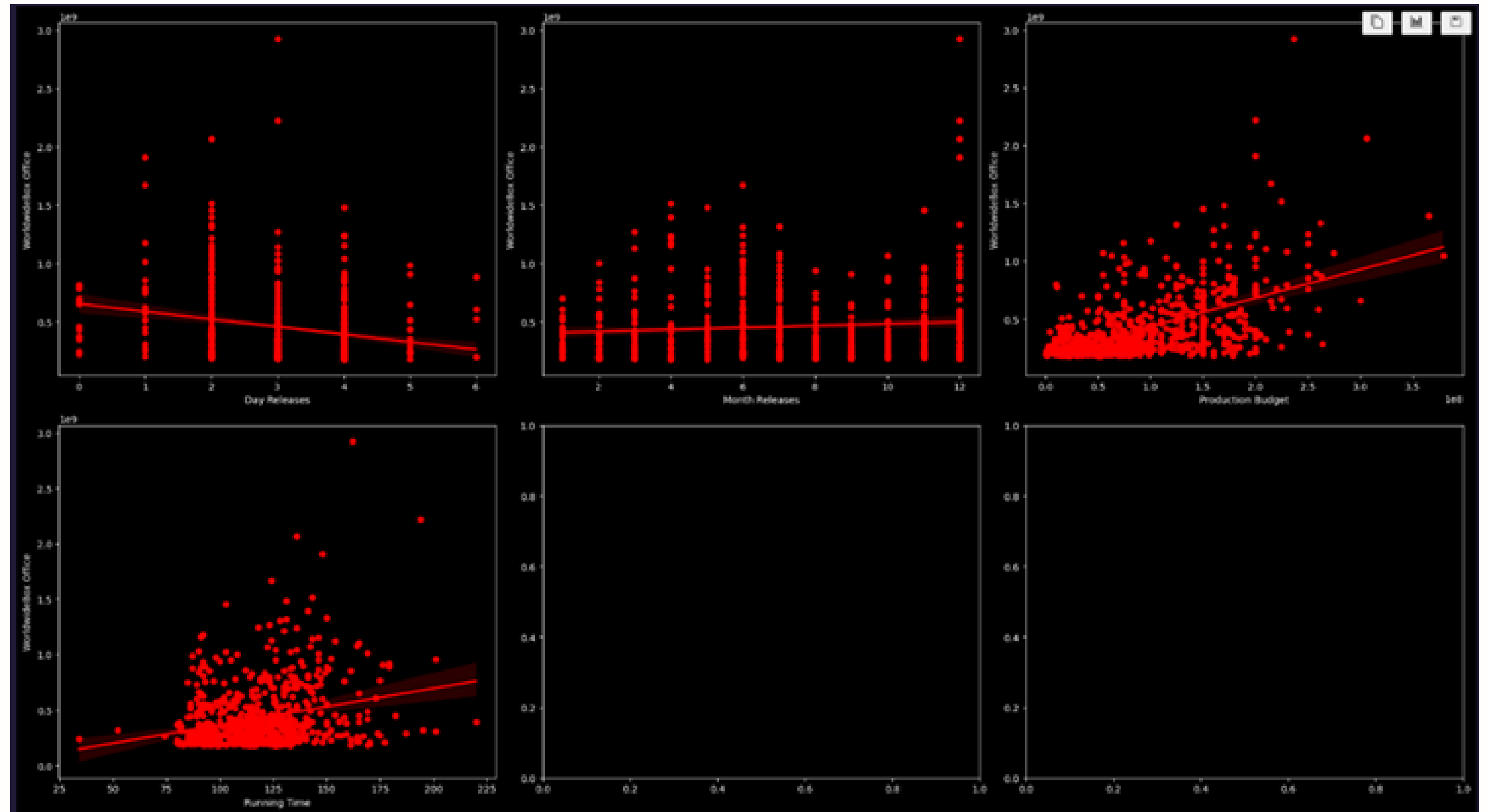
	Production Budget	MPAA	Running Time	Genre	Creative Type	topCompanies	topDirector	Day Releases	Month Releases	holiday_month	origin_isEnglish	topFranchise	countries_isUS
105	580000000.0	7	107.0	0	8	1	1	1.0	2.0	0	1	2	1
68	2580000000.0	4	139.0	1	8	1	1	4.0	6.0	1	1	2	1
479	2100000000.0	4	139.0	0	4	1	1	4.0	6.0	1	1	0	1
399	550000000.0	3	91.0	1	5	1	0	4.0	6.0	1	1	2	1
434	1300000000.0	3	130.0	7	5	1	1	3.0	12.0	1	1	2	1

Dữ liệu huấn luyện sau khi chuẩn hóa dữ liệu

TRÍCH XUẤT ĐẶC TRƯNG

Độ tương quan giữa các đặc trưng với biến mục tiêu

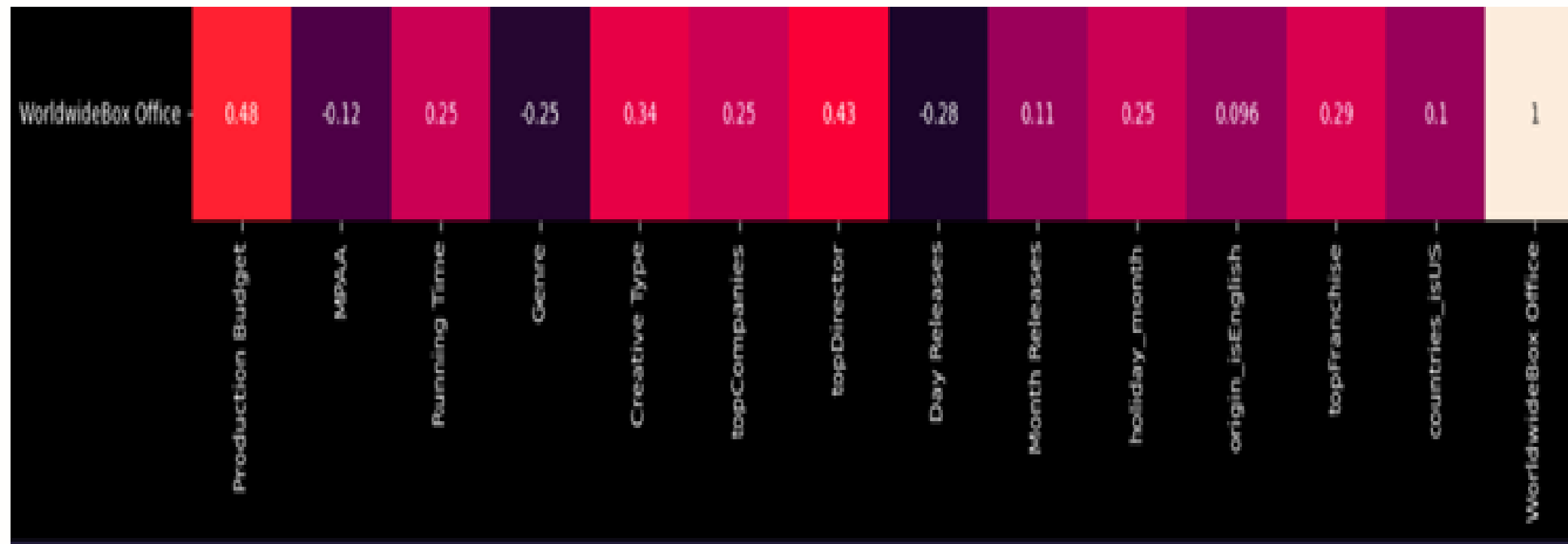
- Tiến hành trực quan hóa sự tương quan của các đặc trưng dạng số so với đặc trưng mục tiêu là ‘WorldwideBox Office’ bằng sơ đồ regplot. Ta thu được sơ đồ dưới đây:



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

TRÍCH XUẤT ĐẶC TRƯNG

- 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 Releases' cũng có sự tương quan nhẹ. Còn đặc trưng 'Month Releases' có độ tương quan không cao.



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
```

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

TRÍCH XUẤT ĐẶC TRƯNG

- 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'.

MÔ HÌNH HÓA DỮ LIỆU

Mô hình Logistic Regression

a. Chia dữ liệu và train mô hình

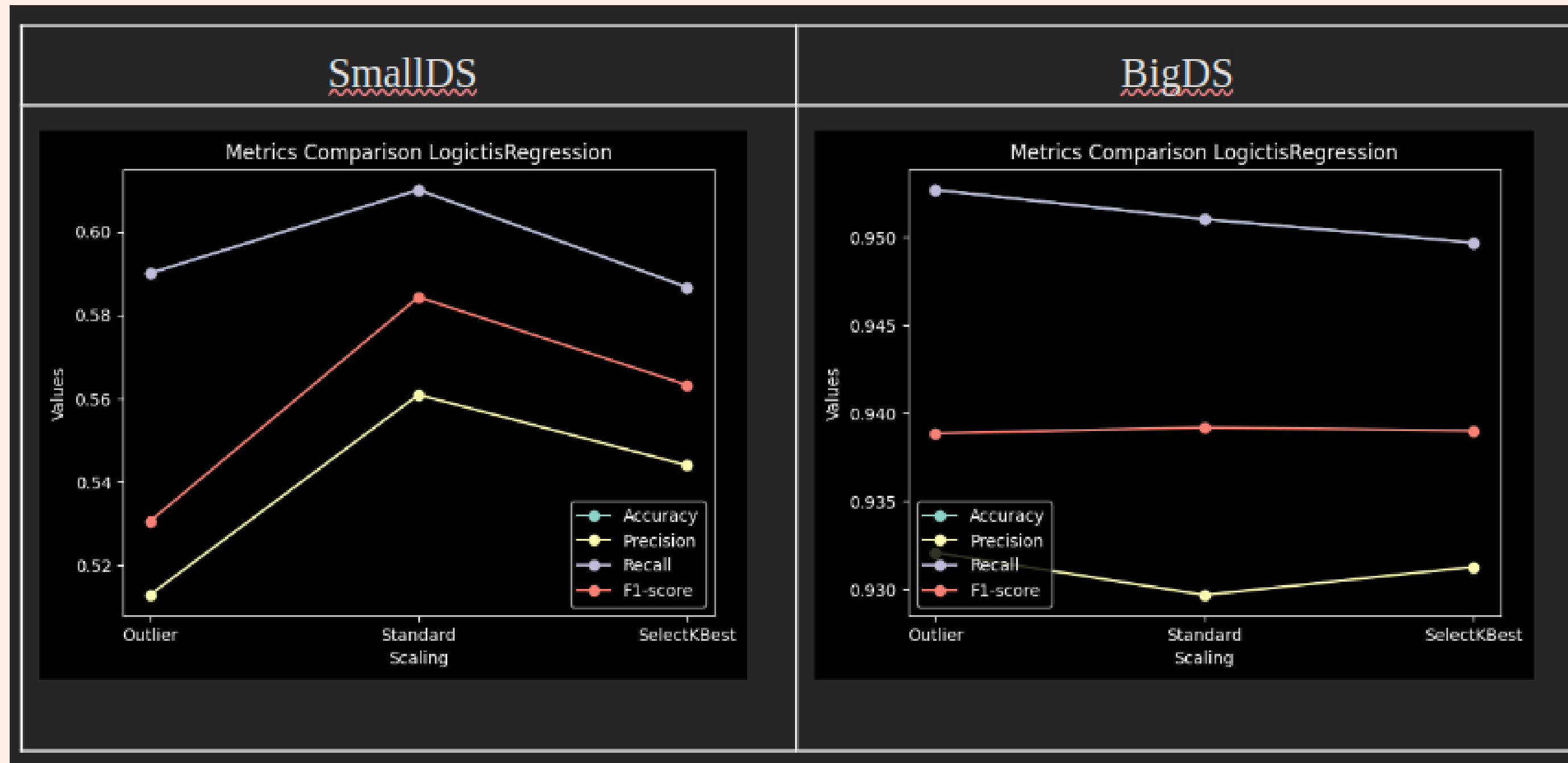
- Bộ tham số của mô hình: Sử dụng thuật toán RandomizedSearchCV để tìm bộ siêu tham số tối ưu cho mô hình.
- Chia tập dữ liệu Huấn luyện/Kiểm thử: theo mô hình 30% kiểm thử.

<u>Tên tham số</u>	<u>Mảng giá trị</u>	<u>Ý nghĩa</u>
C	<u>uniform</u> (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

MÔ HÌNH HÓA DỮ LIỆU

Mô hình Logistic Regression

b. Đồ thị thể hiện hiệu suất

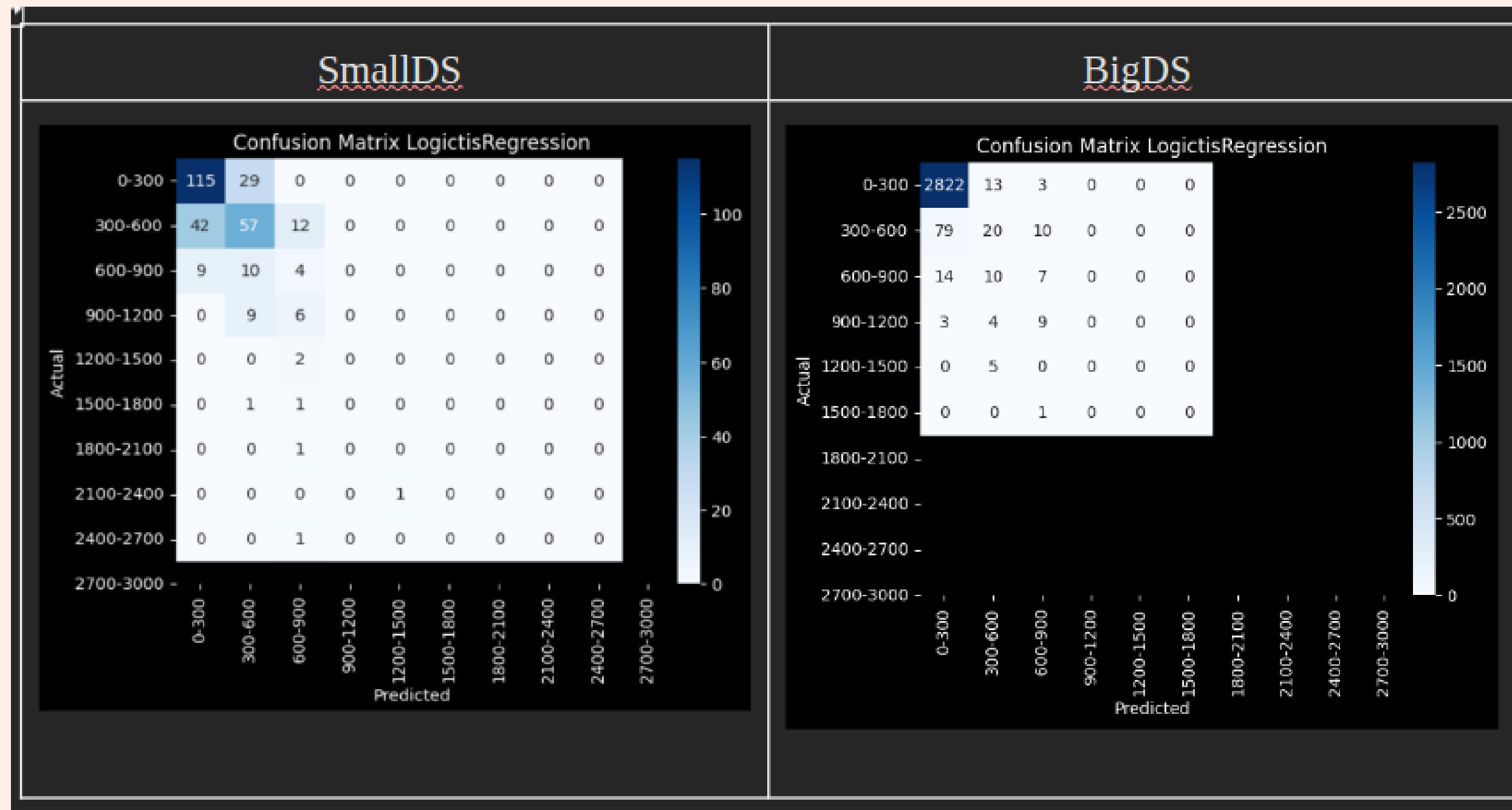


Đồ thị thể hiện hiệu suất 4 metrics trên mô hình logistic regresstion

MÔ HÌNH HÓA DỮ LIỆU

Mô hình Logistic Regression

c. Đồ thị dự đoán so với thực tế



Ma trận nhầm lẫn kết quả dự đoán với kết quả thực của test của mô hình Logistic Regression

MÔ HÌNH HÓA DỮ LIỆU

Mô hình Logistic Regression

d. Đánh giá mô hình theo các metrics là Accuracy, Precision, Recall, F1-Score

<u>SmallDS</u>					<u>BigDS</u>				
	Accuracy	Precision	Recall	F1-score		Accuracy	Precision	Recall	F1-score
Outlier	0.593333	0.513710	0.593333	0.532991	Outlier	0.952667	0.932077	0.952667	0.938840
Standard	0.606667	0.581959	0.606667	0.592468	Standard	0.951000	0.929664	0.951000	0.939178
SelectKBest	0.613333	0.551845	0.613333	0.577135	SelectKBest	0.949667	0.931263	0.949667	0.938991

Đánh giá mô hình Logistic Regresstion theo các metrics là Accuracy, Precesion, Recall, F1-Score

MÔ HÌNH HÓA DỮ LIỆU

Mô hình Random Forest

a. Chia dữ liệu và train mô hình

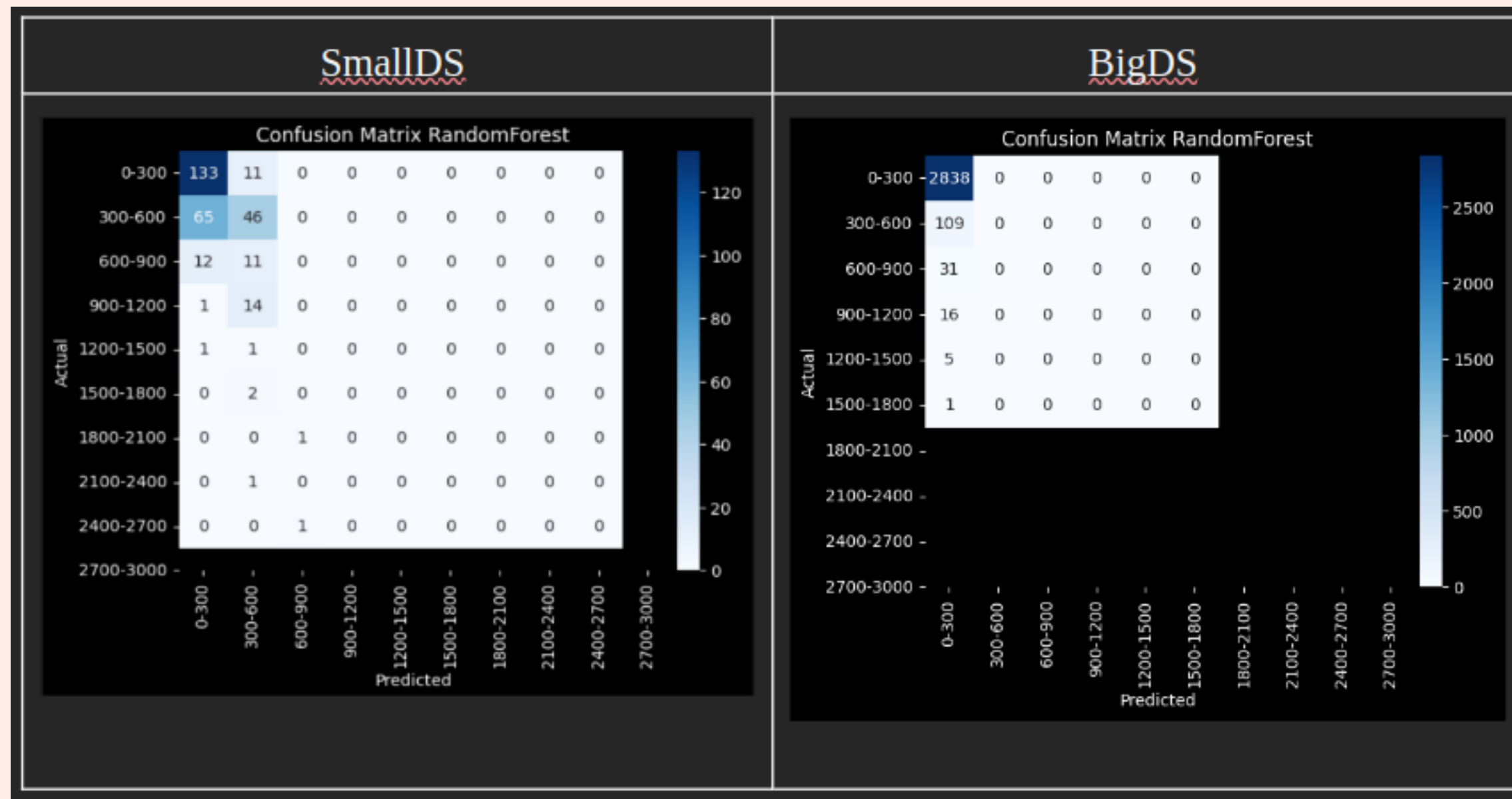
- Bộ tham số của mô hình:
Sử dụng thuật toán RandomizedSearchCV để tìm bộ siêu tham số tối ưu cho mô hình.
- Chia tập dữ liệu Huấn luyện/Kiểm thử: theo mô hình 30% kiểm thử.

<u>Tên tham số</u>	<u>Mảng giá trị</u>	<u>Ý nghĩa</u>
<u>n_estimators</u>	[10, 50, 100]	Số lượng cây trong rừng (forest)
<u>max_depth</u>	[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.
<u>min_samples_split</u>	<u>uniform</u> (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
<u>min_samples_leaf</u>	<u>uniform</u> (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

MÔ HÌNH HÓA DỮ LIỆU

Mô hình Random Forest

b. Đồ thị dự đoán so với thực tế

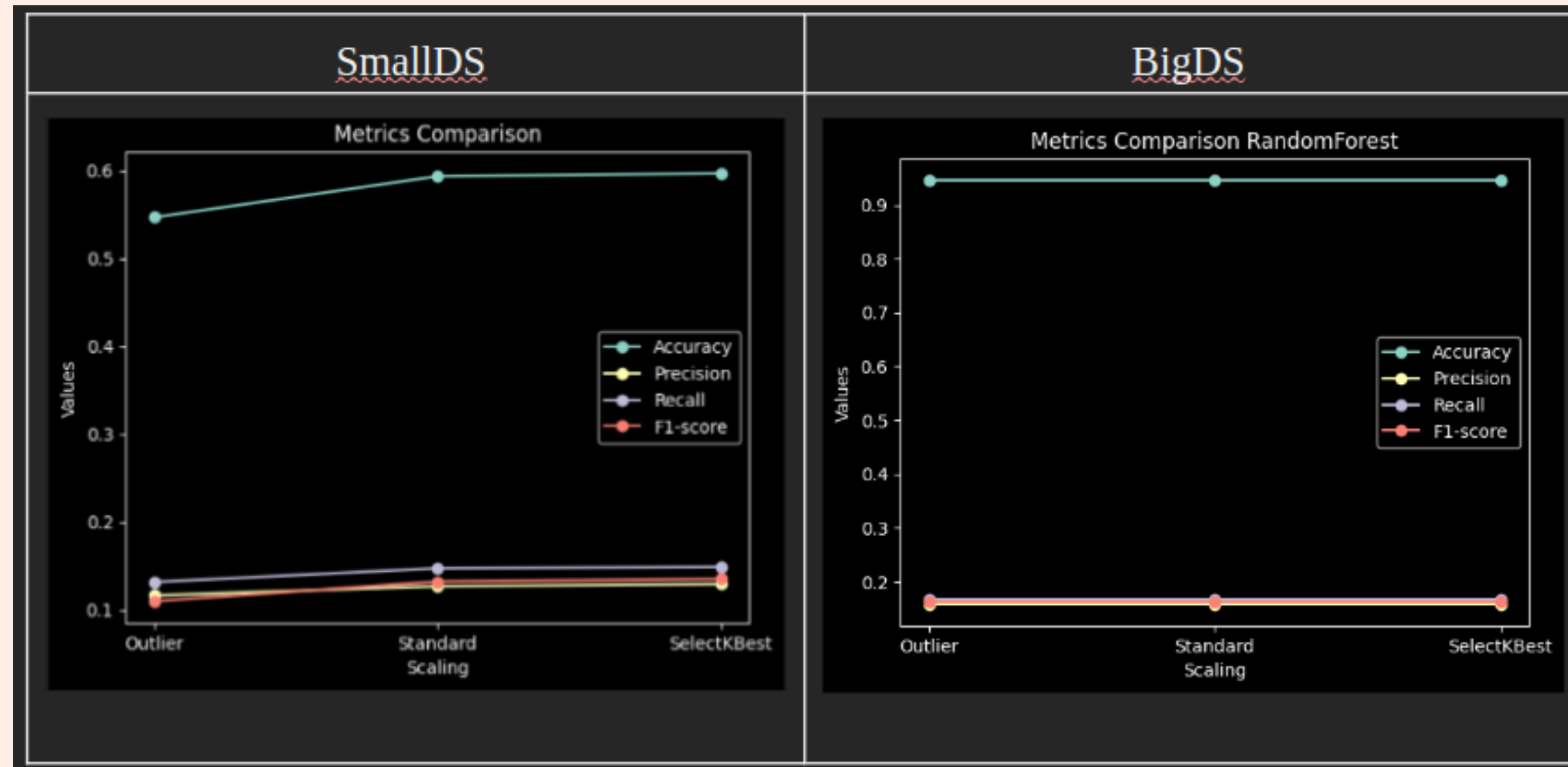


Mã trận nhầm lẫn kết quả dự đoán với kết quả thực của test của mô hình Random Forest

MÔ HÌNH HÓA DỮ LIỆU

Mô hình Random Forest

c. Đồ thị thể hiện hiệu suất



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

MÔ HÌNH HÓA DỮ LIỆU

Mô hình Random Forest

d. Đánh giá mô hình theo các metrics là Accuracy, Precision, Recall, F1-Score

SmallDS					BigDS				
	Accuracy	Precision	Recall	F1-score		Accuracy	Precision	Recall	F1-score
Outlier	0.570000	0.119874	0.139285	0.121892	Outlier	0.946	0.157667	0.166667	0.162042
Standard	0.576667	0.123839	0.140828	0.123154	Standard	0.946	0.157667	0.166667	0.162042
SelectKBest	0.570000	0.119444	0.138597	0.119863	SelectKBest	0.946	0.157667	0.166667	0.162042

Đánh giá mô hình Random Forest theo các metrics là Accuracy, Precision, Recall, F1-Score

KẾT QUẢ VÀ ĐÁNH GIÁ

Kết quả

- Tổng quan tất cả mô hình đã thực nghiệm trên tập kiểm thử, với số liệu thống kê là độ chính xác Accuracy của mô hình sau khi đã lựa chọn K đặc trưng tốt nhất.

Dataset/ Mô hình	Logistic Regression	RandomForestClassifier
SmallDS(1000 samples)	61.33%	57.00%
BigDS(10000 samples)	94.96%	94.60%

Đánh giá

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.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

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).

THANK YOU !