**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**VIỆN TRÍ TUỆ NHÂN TẠO**

# Báo cáo đề tài

# Dự đoán doanh thu phim điện ảnh

**Giảng viên hướng dẫn:** TS. Trần Quốc Long

**Sinh viên thực hiện:** Nguyễn Phương Trang – 22022656

Đỗ Thị Thuỳ Trang – 22022617

Nguyễn Minh Hường – 22022542

**Lớp môn học:** 2324II_INT3405_3

**Github**

*Hà Nội, Tháng 5 năm 2024*

# 1. Giới thiệu

## 1.1 Đặt vấn đề

- Sản xuất và phát hành phim điện ảnh (chiếu rạp) được ví như một ngành công nghiệp "hái ra tiền" vì chỉ cần một bộ phim ăn khách có thể đem lại lợi nhuận lên đến hàng trăm triệu đô la. Vì vậy, mức độ cạnh tranh giữa các nhà sản xuất rất cao, và việc phát hành một bộ phim đòi hỏi sự tính toán cẩn thận trong từng chi tiết, từ kịch bản, đạo diễn, diễn viên, thời điểm phát hành, cách thức quảng bá, cho đến kinh phí đầu tư,...
- Tuy nhiên, đây là một bài toán khó, kể cả với con người, và việc ước tính doanh thu dự kiến của bộ phim đóng vai trò quan trọng trong việc giải quyết bài toán này. Việc phát triển một mô hình AI có khả năng dự đoán doanh thu của một bộ phim chưa phát hành dựa trên các đặc điểm của nó không chỉ giúp dễ dàng ra quyết định đầu tư mà còn mang lại lợi nhuận tối đa cho nhà sản xuất.
- Dựa trên tập dữ liệu, tự crawl từ trang IMDb và lấy từ Kaggel, và bài toán đã nêu ở trên, chúng em đã thử nghiệm với các mô hình khác nhau để dự đoán doanh thu. Ngoài việc đưa ra dự đoán chính xác, chúng em còn quan tâm đến việc nghiên cứu các đặc điểm của phim và sự kết hợp các đặc điểm đó ảnh hưởng như nào lên dự đoán. Chúng em cũng muốn xem hiệu quả của các loại mô hình khác nhau trong bối cảnh của vấn đề.

## 1.2 Mục tiêu

- Dự án này nhằm mục tiêu xây dựng một mô hình dự đoán doanh thu phim dựa trên các thông tin đầu vào như ngân sách sản xuất, ngôn ngữ, thời gian chiếu, thể loại phim, nhà sản xuất, v.v…Mục tiêu cụ thể bao gồm:
    + Thu thập dữ liệu
    + Tiền xử lý dữ liệu
    + Xử lý và trực quan hóa dữ liệu
    + Dùng linear regression và ridge regression làm baseline model
    + Cải tiến model dự đoán dùng bagging, Gradient boosting, random forests, XGBoosting
    + Diễn giải kết quả, ưu nhược điểm của mô hình, sử dụng các phương pháp đánh giá mô hình như $R^2$ score, mean absolute percentage error (MAPE).

+ Chúng em hy vọng rằng với mục tiêu này, dự án sẽ mang lại giá trị thực tế, củng cố, nâng cao kiến thức về machine learning và các kỹ năng mềm khác.

## 1.3 Giới thiệu về tập dữ liệu

### 1.3.1 Tập dữ liệu IMDb

| Các trường trong tập data IMDb | Định nghĩa |
|---|---|
| Domestic Opening | Doanh thu mở màn |
| Name | Tên phim |
| Domestic Distributor | Nhà sản xuất |
| Earliest Release Date | Ngày và khu vực phát hành |
| MPAA | Giới hạn độ tuổi |
| Running Time | Thời lượng phim |
| Genres | Thể loại |
| Overview | Tóm tắt nội dung |
| Actors | Tên các diễn viên |
| Film makers | Tên đạo diễn, kịch bản |
| Worldwide | Doanh thu toàn cầu |
| Domestic | Doanh thu nội địa |
| Foreign | Doanh thu quốc tế |

- Tập dữ liệu có 5456 phim

### 1.3.2 Tập dữ liệu Kaggle

| Các trường trong tập data Kaggle | Định nghĩa |
|---|---|
| id | id |
| belongs_to_collection | Gộp id, title, poster_path và backdrop_path |
| budget | Kinh phí sản xuất |
| genres | Thể loại |
| Homepage | Link đến trang chủ của phim |

| | |
|---|---|
| imdb_id | Id của phim ở trang IMDb |
| original_language | Ngôn ngữ |
| original_title | Tên |
| overview | Tóm tắt nội dung |
| popularity | Chỉ số độ phổ biến |
| poster_path | Link đến poster phim |
| production_companies | Nhà sản xuất |
| production_countries | Quốc gia phát hành |
| release_date | Ngày phát hành |
| runtime | Thời lượng phim |
| spoken_languages | Các ngôn ngữ được dịch ra |
| status | Trạng thái phát hành/chưa phát hành |
| tagline | Hashtag của phim |
| title | Tên |
| Keywords | Từ khoá liên quan đến phim |
| cast | Thông tin các diễn viên |
| crew | Thông tin đạo diễn, biên kịch |
| revenue | Doanh thu phim |

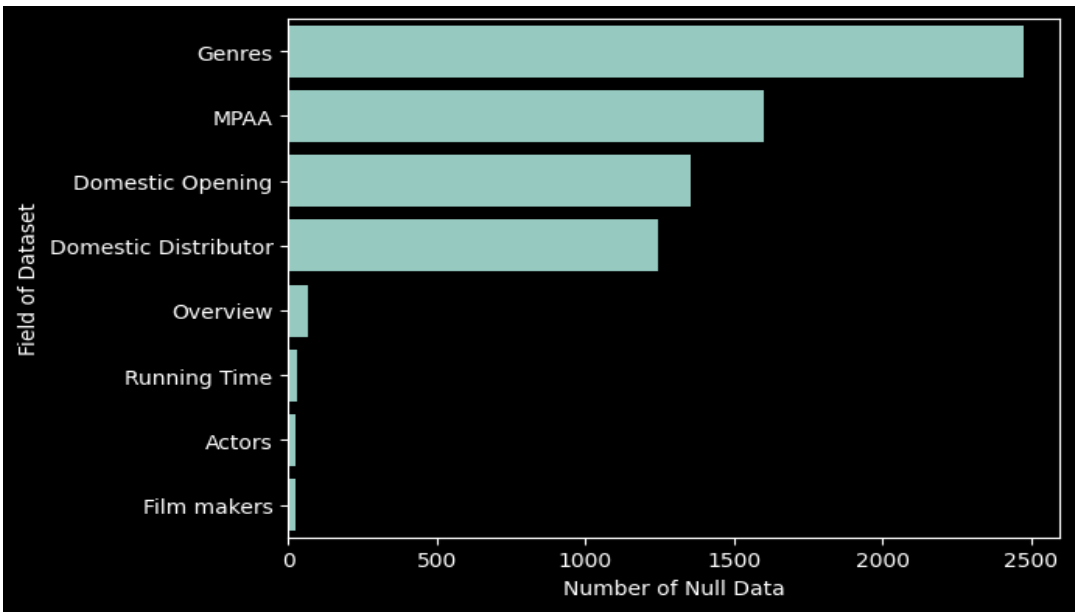
- Tập dữ liệu có 3000 phim

# 2. Tập dữ liệu

## 2.1 Tiền xử lý dữ liệu

### 2.1.1 Tiền xử lý dữ liệu với tập data IMDb

- Đầu tiên, ta xem tổng các giá trị null trong dataset




- Vì lượng null của trường Genres (thể loại) lớn nên ta chỉ còn cách drop các ô null đi.
- Tách trường Earliest Release Date - dạng February 28, 2024 (EMEA, APAC) thành Earliest Release date – dạng 2024-02-28 và Earliest Release region – dạng EMEA, APAC
- Tiếp tục thêm các trường release_day (các ngày trong tuần đánh số từ 0-6), release_month, release_year từ trường Earliest Release date

| | Earliest Release date | release_day | release_month | release_year |
|---|---|---|---|---|
| 0 | 2024-02-28 | 2 | 2 | 2024 |
| 1 | 2024-03-27 | 2 | 3 | 2024 |
| 2 | 2024-03-06 | 2 | 3 | 2024 |
| 3 | 2024-03-20 | 2 | 3 | 2024 |
| 4 | 2024-02-14 | 2 | 2 | 2024 |
| ... | ... | ... | ... | ... |
| 5450 | 2000-09-22 | 4 | 9 | 2000 |
| 5451 | 1985-01-18 | 4 | 1 | 1985 |
| 5453 | 2000-05-12 | 4 | 5 | 2000 |
| 5454 | 2000-09-29 | 4 | 9 | 2000 |
| 5455 | 2000-08-11 | 4 | 8 | 2000 |

- Chuyển trường Running Time – dạng 2 hr 46 min thành trường Running Time (minutes) – dạng 166

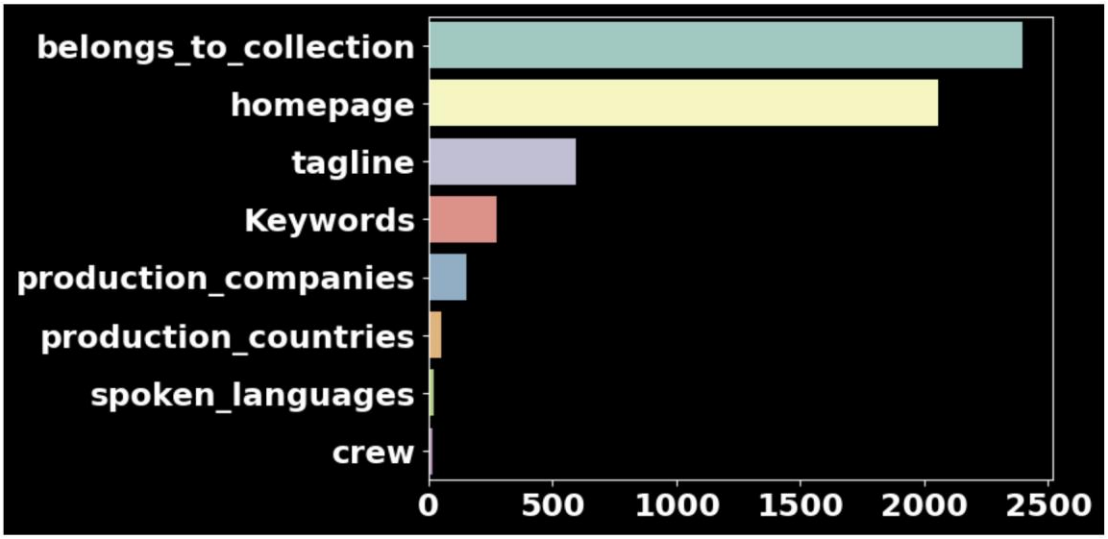
```
0        166
1        115
2         94
3        115
4        107
        ...
5450     100
5451      99
5453     112
5454      94
5455      87
Name: Running Time (minutes), Length: 1753, dtype: int64
```

- Xử lý các hàng bị trùng nhau do quá trình crawl

| | Name | Film makers | Domestic Distributor | MPAA | Genres | Actors | Earliest Release region | Running Time (minutes) | release_day | release_month | release_year |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 78 | Dune | ['Denis Villeneuve', 'Jon Spaihts', 'Denis Vil... | Warner Bros. | PG-13 | Action Adventure Drama Sci-Fi | ['Timothée Chalamet', 'Rebecca Ferguson', 'Zen... | France | 155 | 2 | 9 | 2021 |
| 164 | Dune | ['Denis Villeneuve', 'Jon Spaihts', 'Denis Vil... | Warner Bros. | PG-13 | Action Adventure Drama Sci-Fi | ['Timothée Chalamet', 'Rebecca Ferguson', 'Zen... | France | 155 | 2 | 9 | 2021 |
| 304 | Toy Story | ['John Lasseter', 'John Lasseter', 'Pete Docte... | Walt Disney Studios Motion Pictures | G | Adventure Animation Comedy Family Fantasy | ['Tom Hanks', 'Tim Allen', 'Don Rickles', 'Jim... | Domestic | 81 | 2 | 11 | 1995 |
| 564 | Spider-Man: No Way Home | ['Jon Watts', 'Chris McKenna', 'Erik Sommers',... | Sony Pictures Releasing | PG-13 | Action Adventure Fantasy Sci-Fi | ['Tom Holland', 'Zendaya', 'Benedict Cumberbat... | 14 markets | 148 | 2 | 12 | 2021 |
| 589 | Spider-Man: No Way Home | ['Jon Watts', 'Chris McKenna', 'Erik Sommers',... | Sony Pictures Releasing | PG-13 | Action Adventure Fantasy Sci-Fi | ['Tom Holland', 'Zendaya', 'Benedict Cumberbat... | 14 markets | 148 | 2 | 12 | 2021 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4836 | The Barbarian Invasions | ['Denys Arcand', 'Denys Arcand', 'Daniel Louis... | Alliance Atlantis Vivafilm | R | Comedy Crime Drama Mystery Romance | ['Rémy Girard', 'Dorothée Berryman', 'Stéphane... | Domestic | 99 | 4 | 5 | 2003 |
| 4842 | Winged Migration | ['Jacques Perrin', 'Jacques Cluzaud', 'Michel ... | Mongrel Media | G | Documentary | ['Jacques Perrin', 'Philippe Labro'] | France | 98 | 2 | 12 | 2001 |

2.1.2 Tiền xử lý dữ liệu với tập data Kaggle

- Như với dataset trên, ta cũng xem các trường null

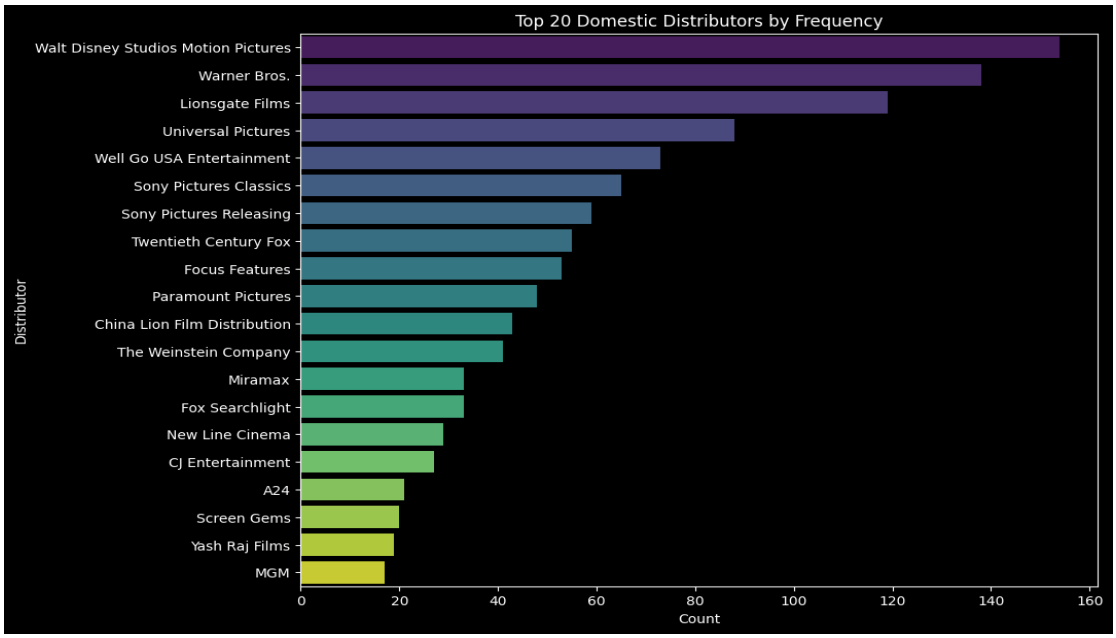


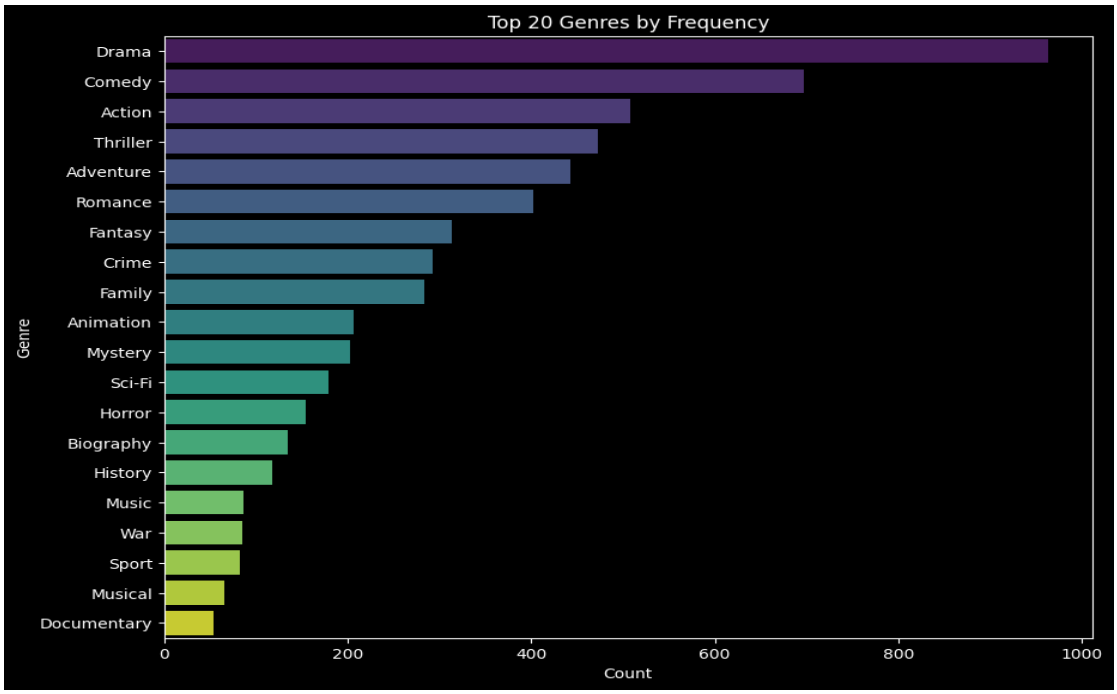
- Ta sẽ drop các ô có giá trị null đi

## 2.2 Xử lý và trực quan hoá dữ liệu

### 2.2.1 Xử lý và trực quan hoá dữ liệu với dataset IMDb

- Với trường Domestic Distributor (nhà sản xuất) và Genres (thể loại), ta sẽ xem các nhà sản xuất tiêu biểu.

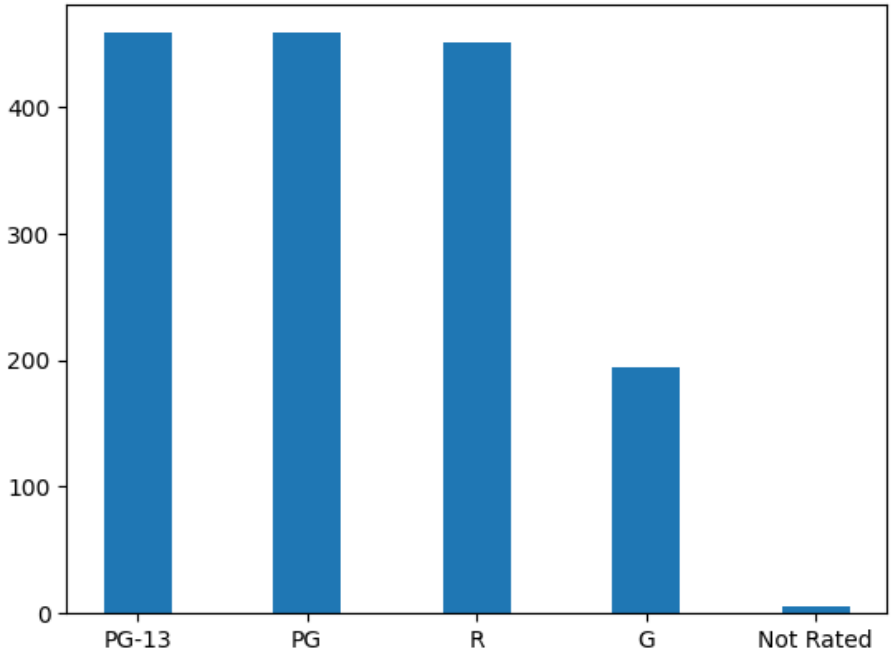


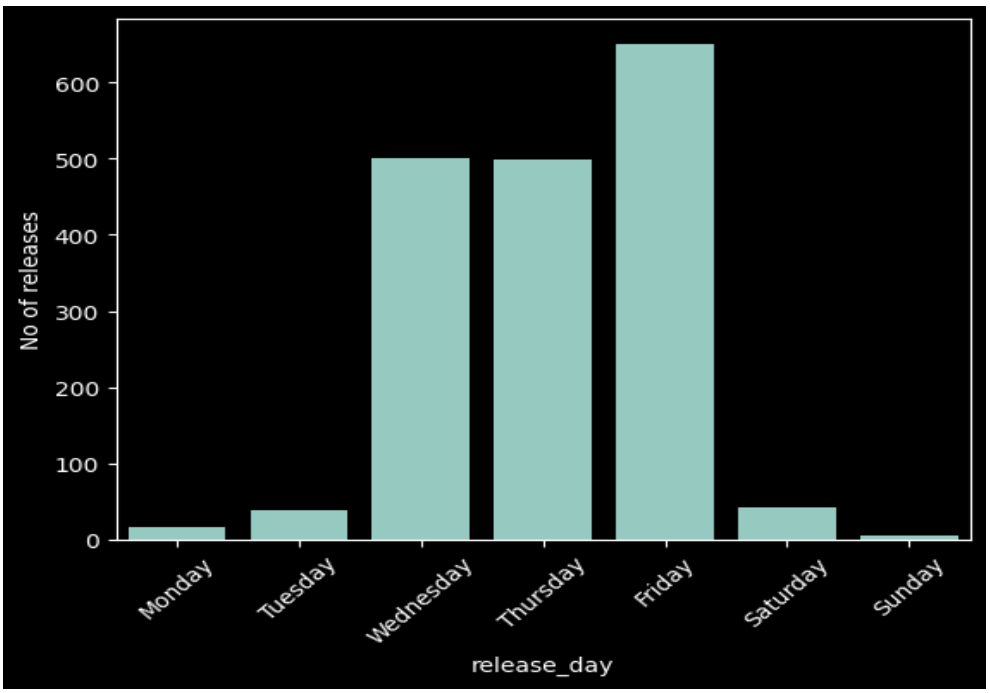

- Mỗi phim thuộc nhiều thể loại khác nhau, để thuận lợi cho việc tính toán của model ta sẽ one-hot encoding trường Genres và cả Domestic Distributor (nhà sản xuất) về dạng như sau:

| | Name | Domestic Distributor etc | Walt Disney Studios Motion Pictures | Warner Bros. | Lionsgate Films | Universal Pictures | Well Go USA Entertainment | Sony Pictures Classics | Sony Pictures Releasing | Focus Features | ... | China Lion Film Distribution | Twentieth Century Fox | The Weinstein Company | Er |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dune: Part Two | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | |
| | Godzilla x Kong: The New Empire | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | |
| | Kung Fu Panda 4 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | |
| | Ghostbusters: Frozen Empire | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | ... | 0 | 0 | 0 | |
| | Bob Marley: One Love | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | |
| | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| | Shower | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | ... | 0 | 0 | 0 | |
| | Left Luggage | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | |
| | Hamlet | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | |
| | The Broken Hearts Club: A Romantic Comedy | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | ... | 0 | 0 | 0 | |
| | Cecil B. Demented | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | |

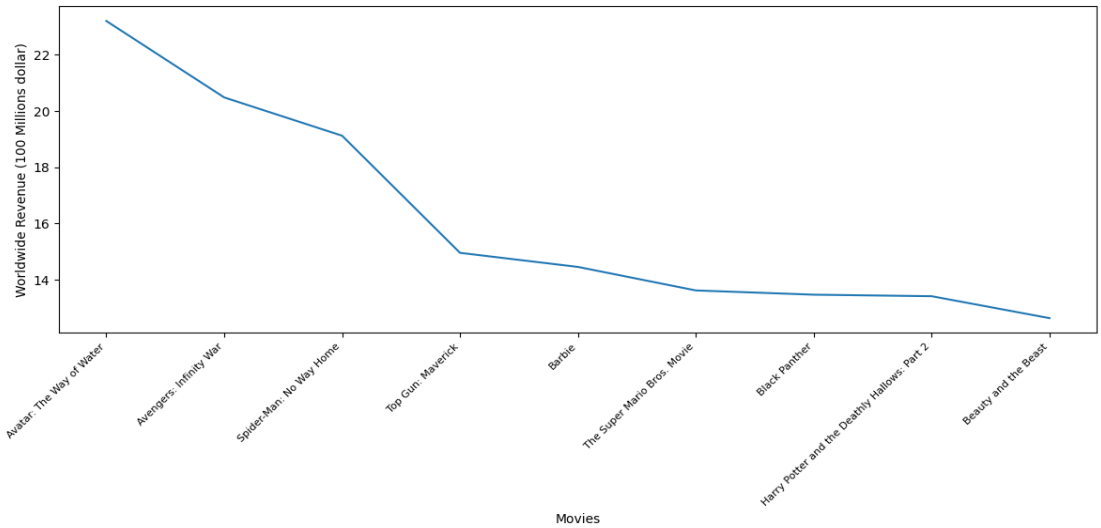
- Giới hạn độ tuổi MPAA được quy ước như sau: G (General Audiences), PG (Parental Guidance Suggested), PG-13 (Parents Strongly Cautioned), R (Restricted)




- Số phim phát hành theo các ngày




- Với trường Name, ta sẽ xử lý bằng cách đếm số từ và số ký tự trong tên phim. Ta thêm 2 trường original_title_letter_count và original_title_word_count.
- Ngoài ra ta cũng normalize 2 trường trên và trường Running time (minutes) bằng cách gọi hàm normalize từ sklearn.preprocessing
- Top 10 phim có doanh thu cao nhất (từ 2000 – 2024)



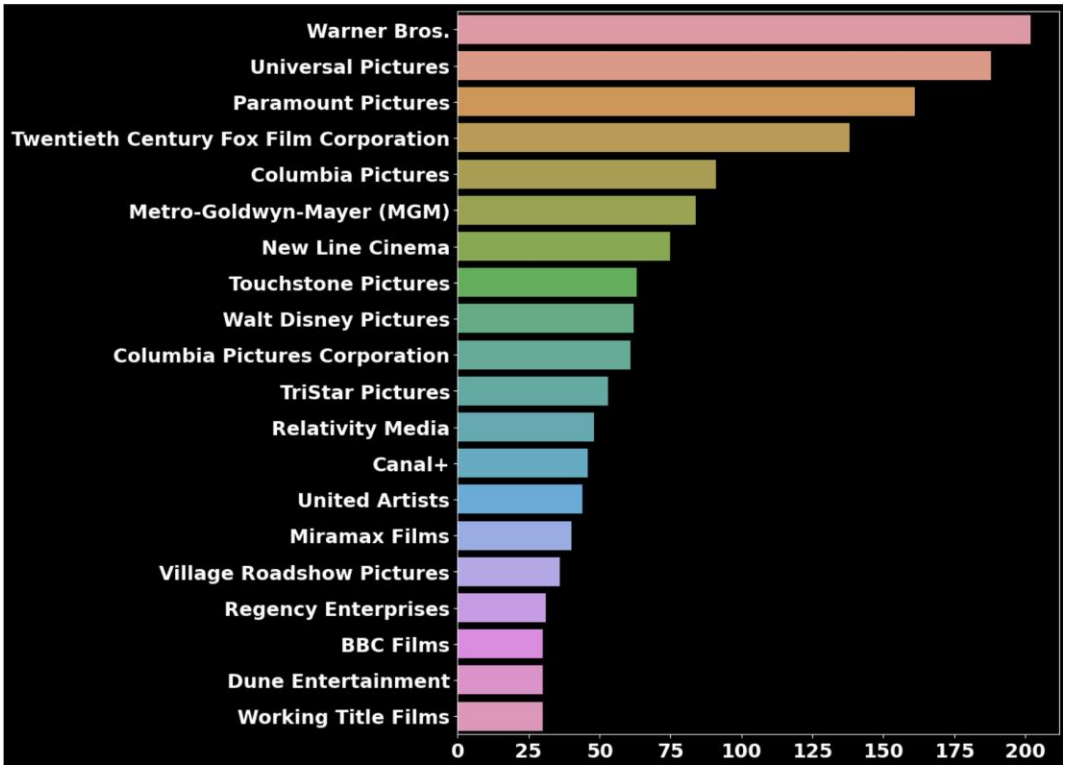

- Ta thấy quá nử số phim này đều có doanh thu trên 1 tỷ đô, điều này sẽ ảnh hưởng đến độ chính xác của model nên ta sẽ drop chúng đi.
- Các trường còn lại Domestic Distributor, Genres (đã one-hot encoding), Film makers, Actors, MPAA, Earliest Release region, Domestic ta sẽ drop đi để thành input cho các model.
- Ta sẽ lấy log của doanh thu, trường Worldwide, để dễ dàng tính toán cho model. Cuối cùng data có dạng x gồm các trường ['Running Time (minutes)',

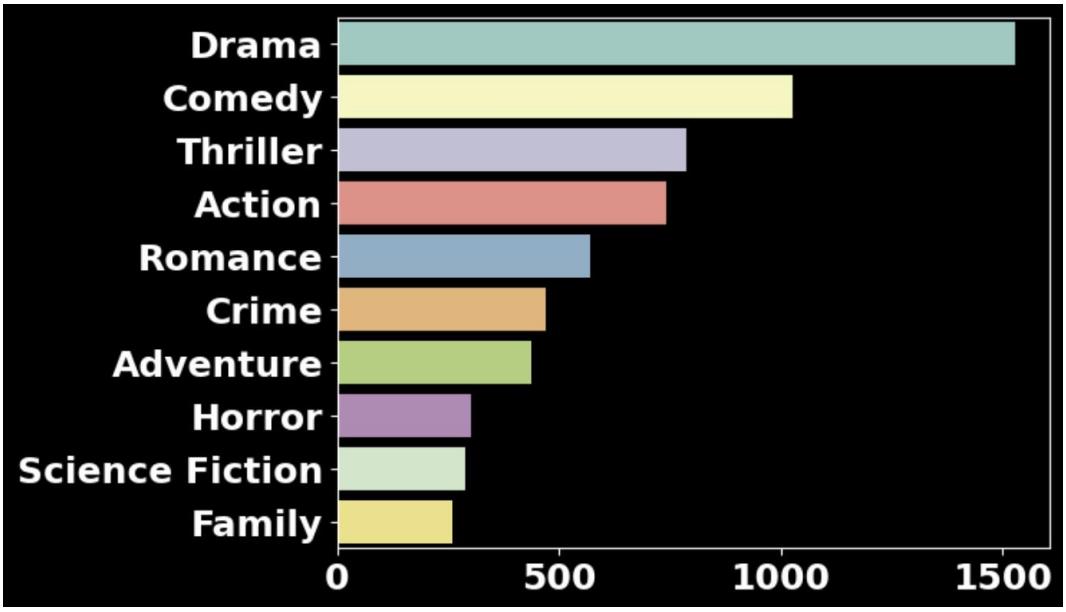
'release_day', 'release_month', 'release_year', 'Domestic Distributor etc', 'Walt Disney Studios Motion Pictures', 'Warner Bros.', 'Lionsgate Films', 'Universal Pictures', 'Well Go USA Entertainment', 'Sony Pictures Classics', 'Sony Pictures Releasing', 'Focus Features', 'Paramount Pictures', 'China Lion Film Distribution', 'Twentieth Century Fox', 'The Weinstein Company', 'CJ Entertainment', 'Miramax', 'Fox Searchlight', 'A24', 'New Line Cinema', 'Yash Raj Films', 'Fathom Events', 'Screen Gems', 'Genres etc', 'Drama', 'Comedy', 'Action', 'Thriller', 'Adventure', 'Romance', 'Fantasy', 'Crime', 'Family', 'Mystery', 'Animation', 'Sci-Fi', 'Horror', 'Biography', 'History', 'Music', 'War', 'Sport', 'Musical', 'Documentary', 'original_title_letter_count', 'original_title_word_count'] và y là log của doanh thu phim

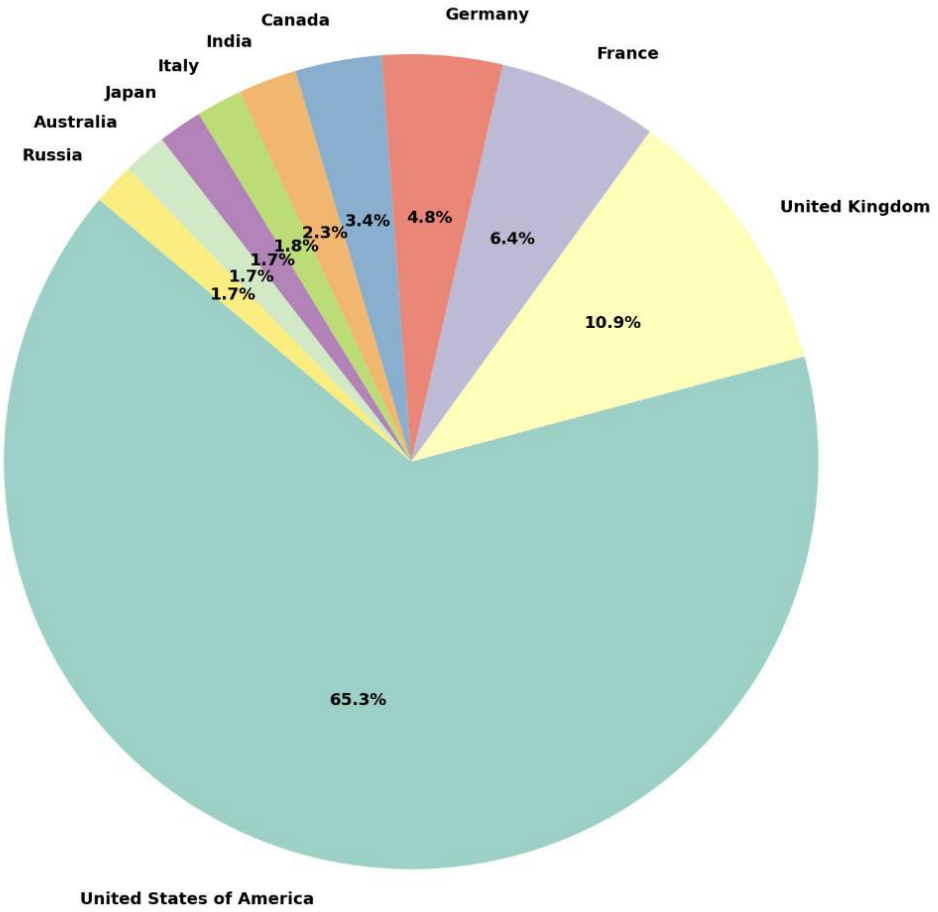
2.2.2 Xử lý và trực quan hoá dữ liệu với dataset Kaggle

- Đầu tiên, ta sẽ xem các keyword phổ biến

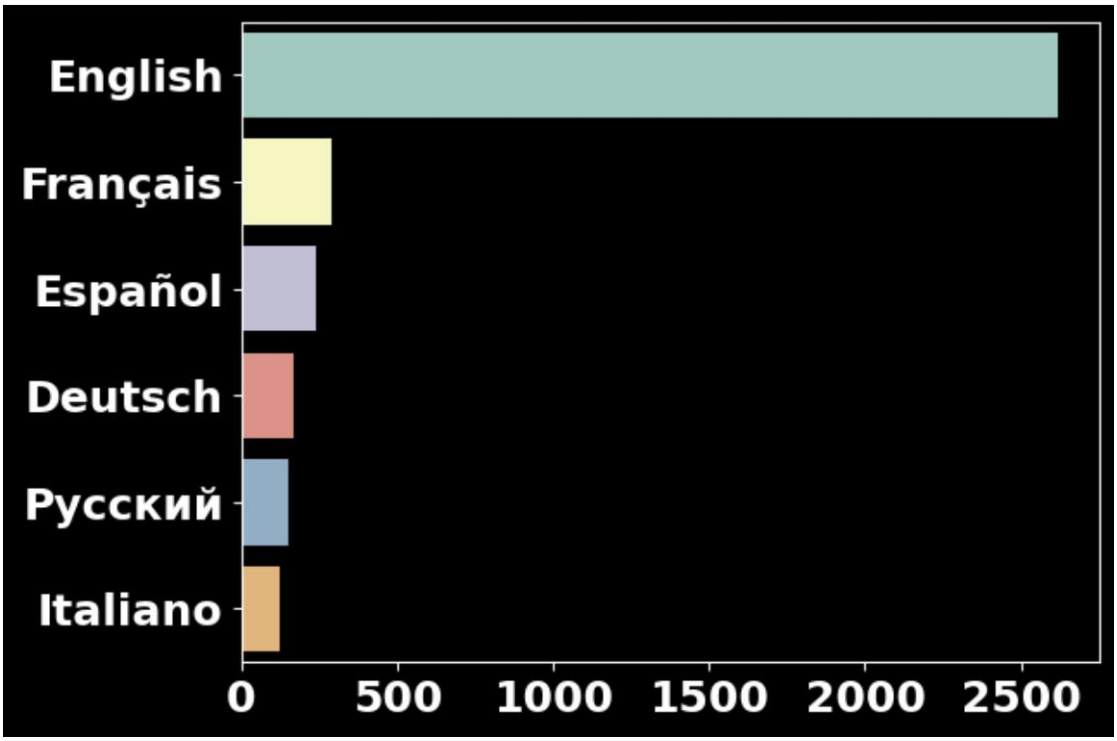

- Tương tự như trên, ta xem các nhà sản xuất và thể loại tiêu biểu nhất. Ngoài ra ở dataset này ta có thêm production_countries.


Warner Bros.
Universal Pictures
Paramount Pictures
Twentieth Century Fox Film Corporation
Columbia Pictures
Metro-Goldwyn-Mayer (MGM)
New Line Cinema
Touchstone Pictures
Walt Disney Pictures
Columbia Pictures Corporation
TriStar Pictures
Relativity Media
Canal+
United Artists
Miramax Films
Village Roadshow Pictures
Regency Enterprises
BBC Films
Dune Entertainment
Working Title Films
0
25
50
75
100
125
150
175
200

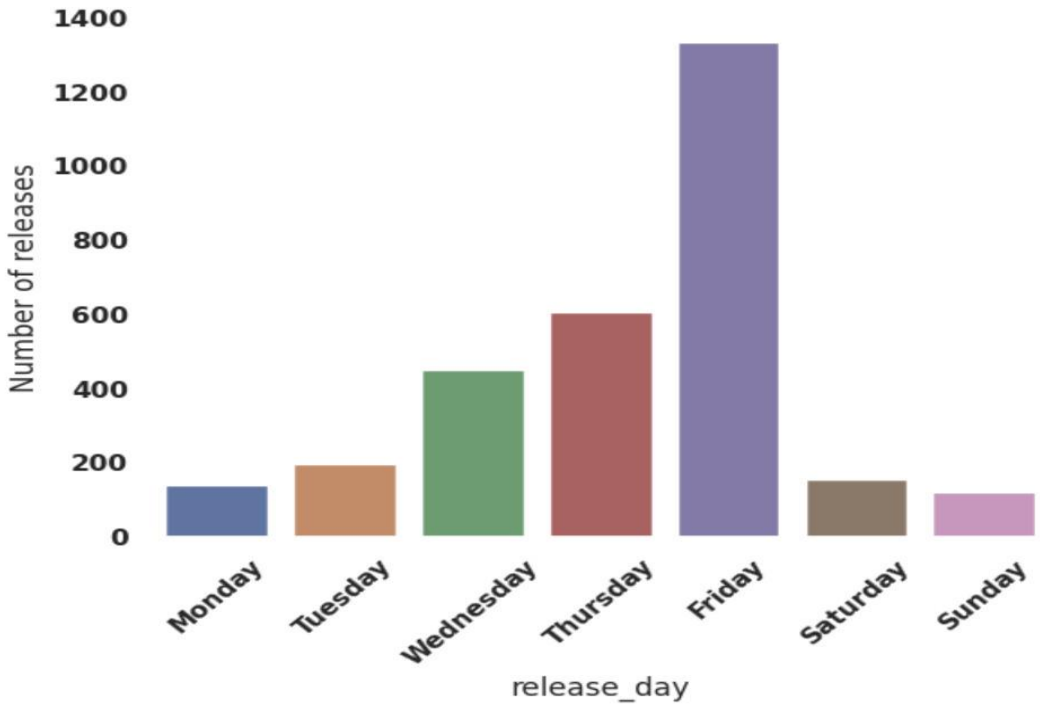

Drama
Comedy
Thriller
Action
Romance
Crime
Adventure
Horror
Science Fiction
Family
0
500
1000
1500



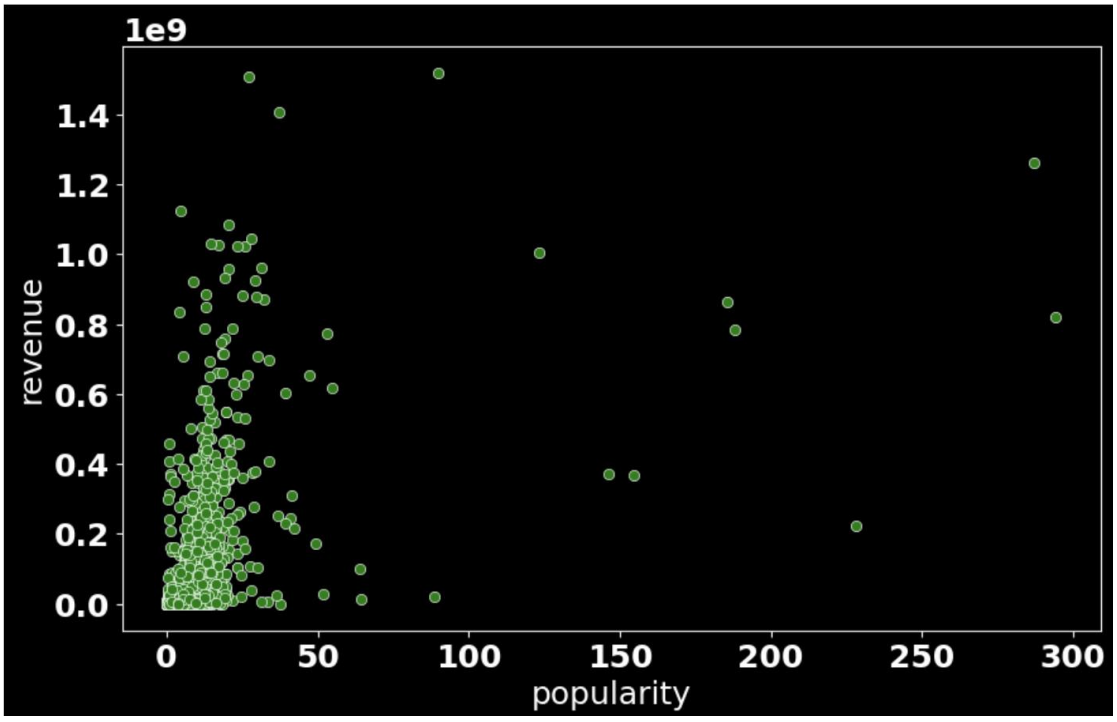

- Ta cũng sẽ one-hot encoding trường production_companies, genres và cả production_countries.
- Các ngôn ngữ phổ biến



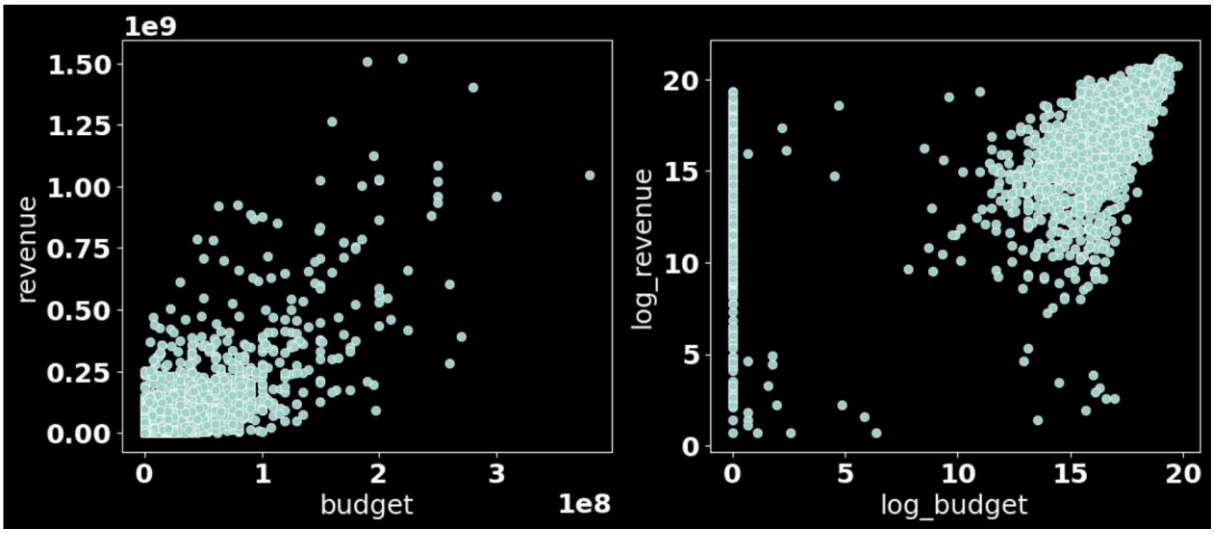
- Ta cũng chia trường release_date thành từng ngày, tháng và năm cụ thể. Biểu đồ dưới là số phim phát hành theo từng ngày.



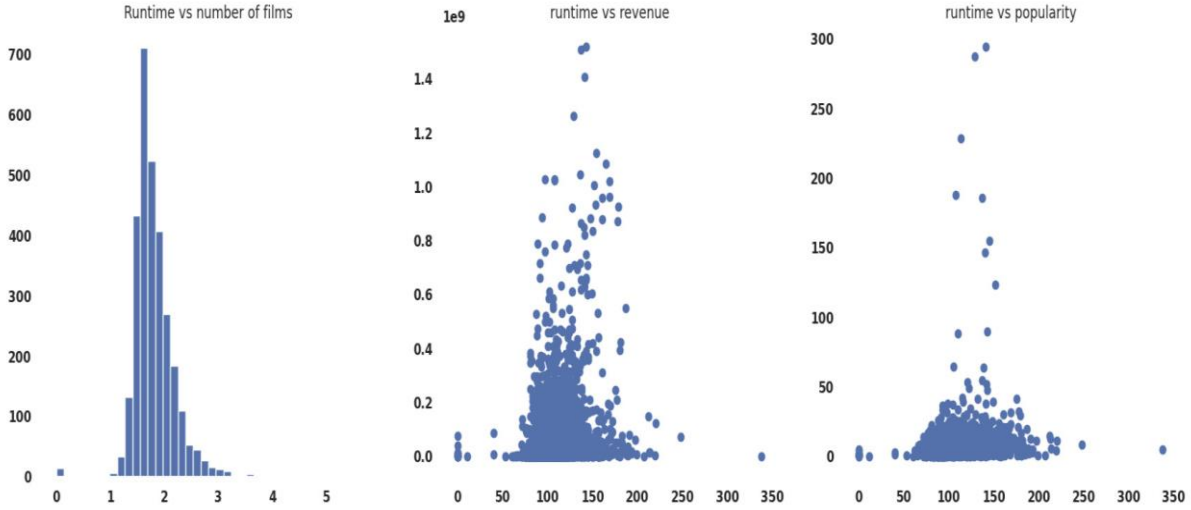

- Độ tương quan giữa chỉ số phổ biến (popularity) với revenue



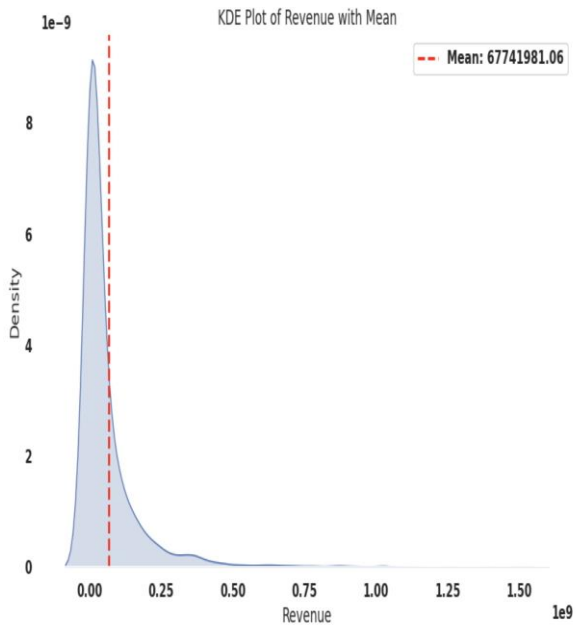

- Với trường budget, ta xem độ tương quan giữa budget và revenue

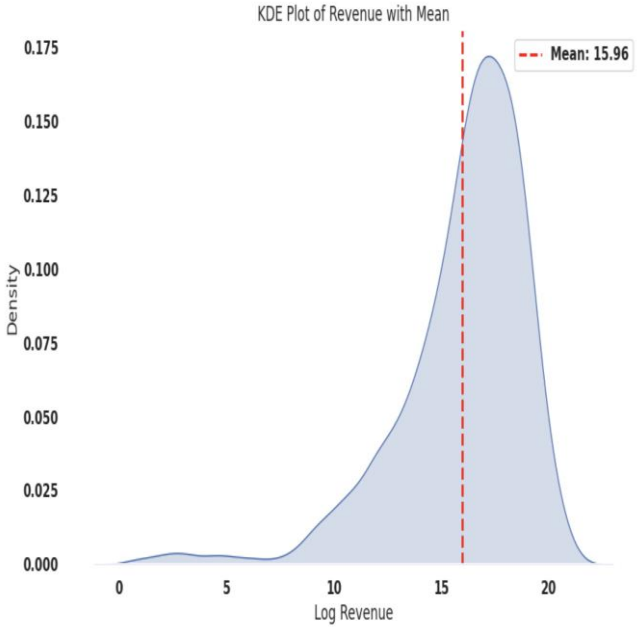



- Thời lượng phim (runtime) với revenue, popularity




- Với trường coi là output của model, ta sẽ xem phân bố của revenue



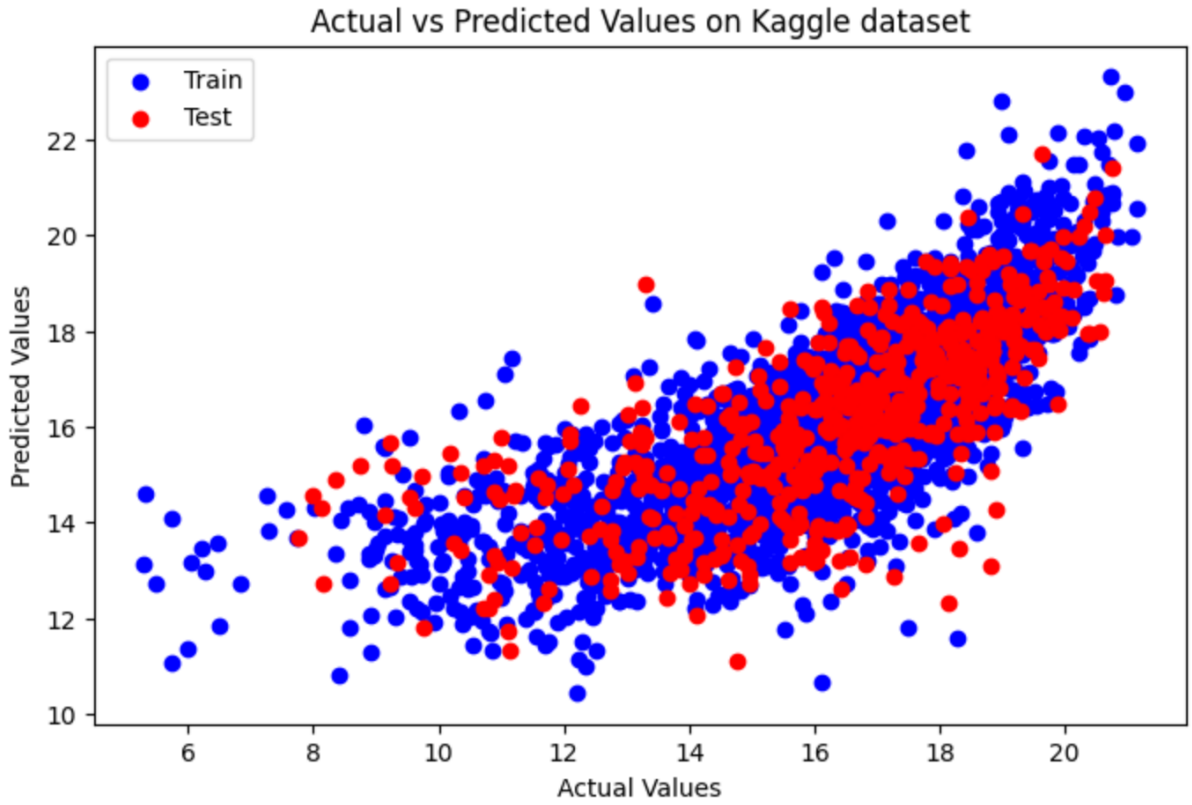
- Ta thấy rằng việc lấy log của doanh thu thì phân bố các giá trị sẽ xung quanh mean và dễ dàng cho việc tính toán hơn nên ta sẽ lấy log của revenue là output của model. Nói cách khác model sẽ dự đoán log của doanh thu phim.
- Ta sẽ drop các trường không dùng tới làm input gồm: belongs_to_collection, homepage, imdb_id, original_title, overview, poster_path, status, tagline, original_language
- Hàm prepare_data để xử lý data trước khi làm việc với model. Trong hàm này, các cột sau được normalize bằng hàm normalie trong sklearn: 'runtime','popularity','budget','_budget_runtime_ratio','_budget_year_ratio','_budget_popularity_ratio','_releaseYear_popularity_ratio', '_releaseYear_popularity_ratio2','_num_Keywords','_num_cast','no_spoken_languages','original_title_letter_count','original_title_word_count', 'title_word_count','overview_word_count','production_countries_count','production_companies_count','cast_count','crew_count', 'genders_0_crew','genders_1_crew','genders_2_crew'
- Sau khi drop, normalize và tính thêm các trường mới thì input đầu vào cho model gồm 168 trường: budget, popularity, runtime, release_day, release_month, release_year, _budget_runtime_ratio, _budget_popularity_ratio, _buget_year_ratio, _releaseYear_popularity_ratio, _releaseYear_popularity_ratio2, has_homepage, _num_Keywords, _num_cast, ....

# 3. Huấn luyện mô hình và diễn giải kết quả

## 3.1 Baseline (Linear Regression)

Hồi quy tuyến tính (Linear Regression) là 1 phương pháp thống kê biểu diễn mối quan hệ giữa 1 biến phụ thuộc vào một hoặc nhiều biến độc lập. Kết quả sau khi chạy model trên Kaggle dataset như sau:

```
Final Training Accuracy: 57.39%
Final Test Accuracy: 48.67%
Train MAPE: 0.09%
Test MAPE: 0.10%
Train MSE: 2.9418734097889705
Test MSE: 3.031450195771525
```

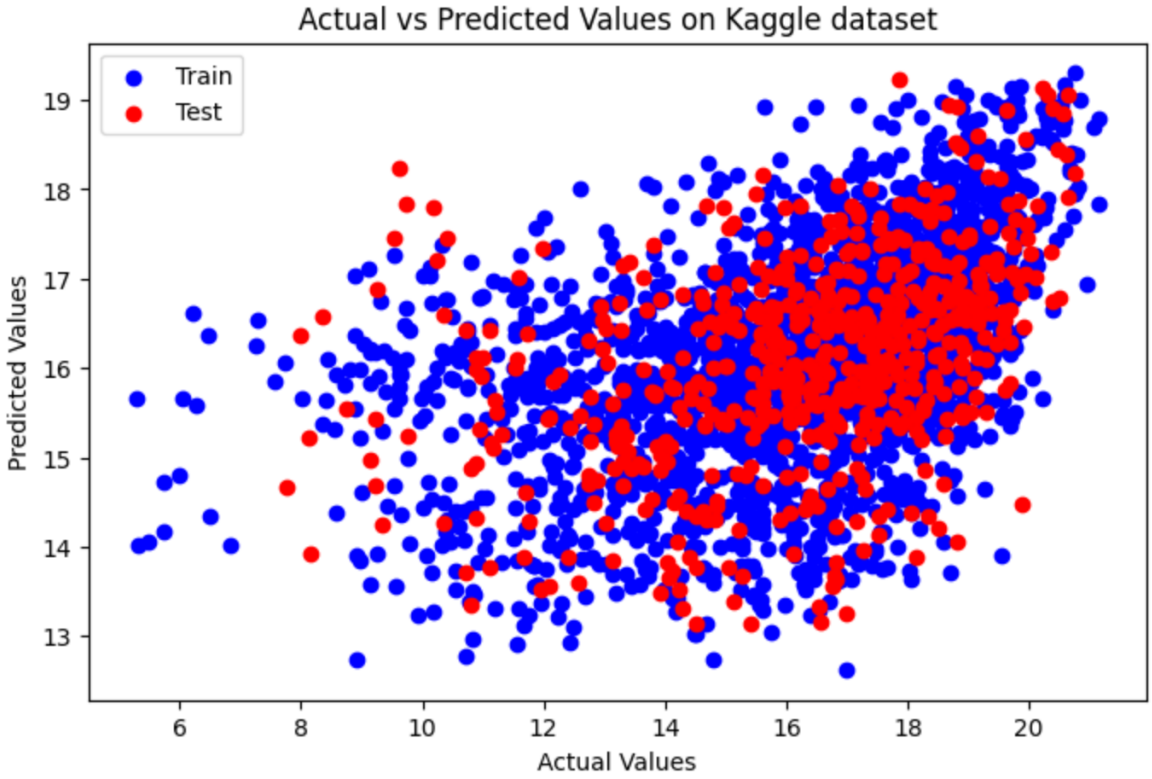



Từ biểu đồ, ta có thể quan sát thấy chỉ số MSE vẫn còn rất cao. Vậy nên, chúng ta sẽ tiếp tục sử dụng PCA để thử tối ưu model.

## 3.2 Linear Regression PCA

PCA (Principal Component Analysis) là một kỹ thuật giảm chiều dữ liệu. Mục đích của PCA là giảm số lượng biến số (dimensions) trong một tập dữ liệu trong khi vẫn giữ lại càng nhiều thông tin quan trọng càng tốt. Kết quả khi kết hợp PCA với Linear Regression như sau:

```
Final Training Accuracy: 20.14%
Final Test Accuracy: 16.05%
Train MAPE: 0.13%
Test MAPE: 0.13%
Train MSE: 5.513590405904071
Test MSE: 5.567107121830867
```




Kết quả cho thấy MSE tăng lên

### 3.3 Bagging

Bagging giúp giảm phương sai và ngăn overfitting, đặc biệt hiệu quả với các mô hình có phương sai cao như cây quyết định (decision trees).

Cách hoạt động của Bagging

Bootstrapping: Tạo nhiều tập dữ liệu con từ tập dữ liệu gốc bằng cách lấy mẫu ngẫu nhiên có hoàn lại. Mỗi tập dữ liệu con (bootstrap sample) có kích thước bằng tập dữ liệu gốc nhưng có thể chứa các mẫu trùng lặp.

Training: Huấn luyện một mô hình riêng biệt (base estimator) trên mỗi tập dữ liệu con. Các mô hình này thường cùng loại, ví dụ như các cây quyết định.

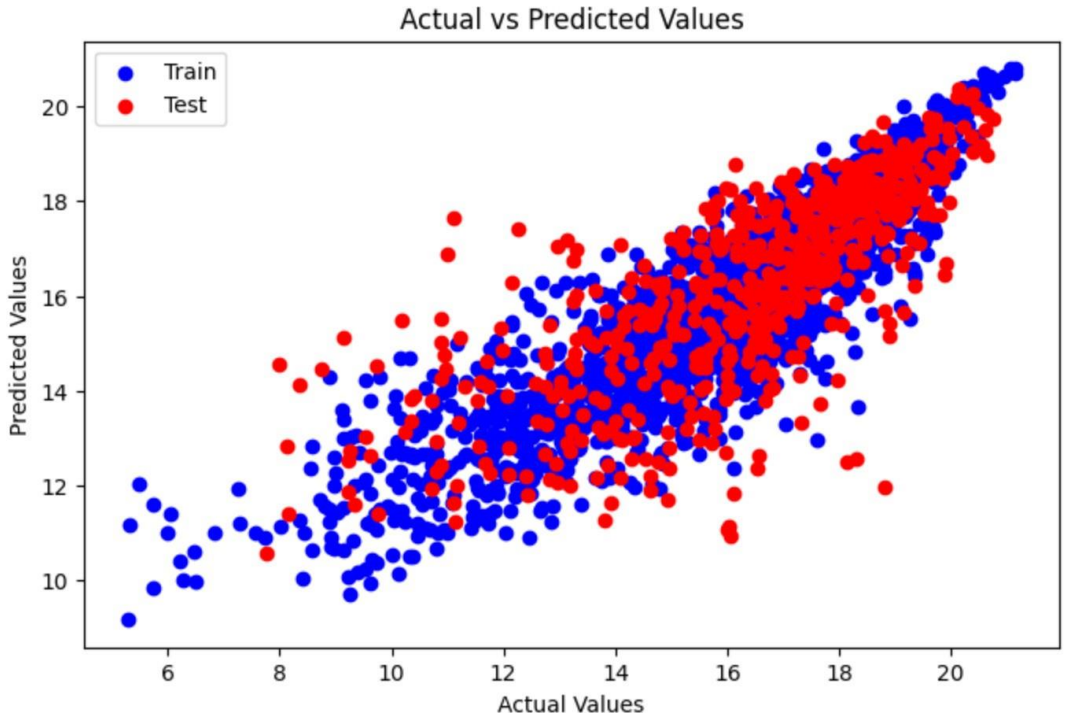
Aggregation: Kết hợp các dự đoán từ tất cả các mô hình đã huấn luyện để tạo ra dự đoán cuối cùng. Đối với bài toán hồi quy, các dự đoán được tính trung bình. Đối với bài toán phân loại, đa số phiếu quyết định lớp cuối cùng.

Sau đây, nhóm em sẽ áp dụng kỹ thuật bagging với model decision tree.

3.4 Decision trees

Cây quyết định hoạt động bằng cách chia tập dữ liệu thành các tập con dựa trên các điều kiện trên các thuộc tính. Quá trình này tiếp tục lặp đi lặp lại đệ quy cho đến khi đạt được các nút lá hoặc khi không thể chia nhỏ thêm nữa. Mỗi lần chia nhỏ, cây sẽ cố gắng giảm thiểu độ nhiễu và tăng độ thuần khiết của các tập con. Chúng ta sẽ xây dựng model decision tree với độ sâu là 10 nút kết hợp với kĩ thuật bagging để cải thiện độ chính xác. Kết quả sau khi thực hiện như sau:

```
Final Training Accuracy: 79.62%
Final Test Accuracy: 54.29%
Train MAPE: 0.06%
Test MAPE: 0.09%
Train MSE: 1.406962251426828
Test MSE: 3.031450195771525
```




3.5 Gradient boosting

Gradient Boosting là một phương pháp tăng cường (boosting) mà mục tiêu chính là kết hợp nhiều mô hình yếu (weak learners), thường là các cây quyết định, để tạo ra một mô hình mạnh hơn. Kết quả của nhóm em sau khi chạy model

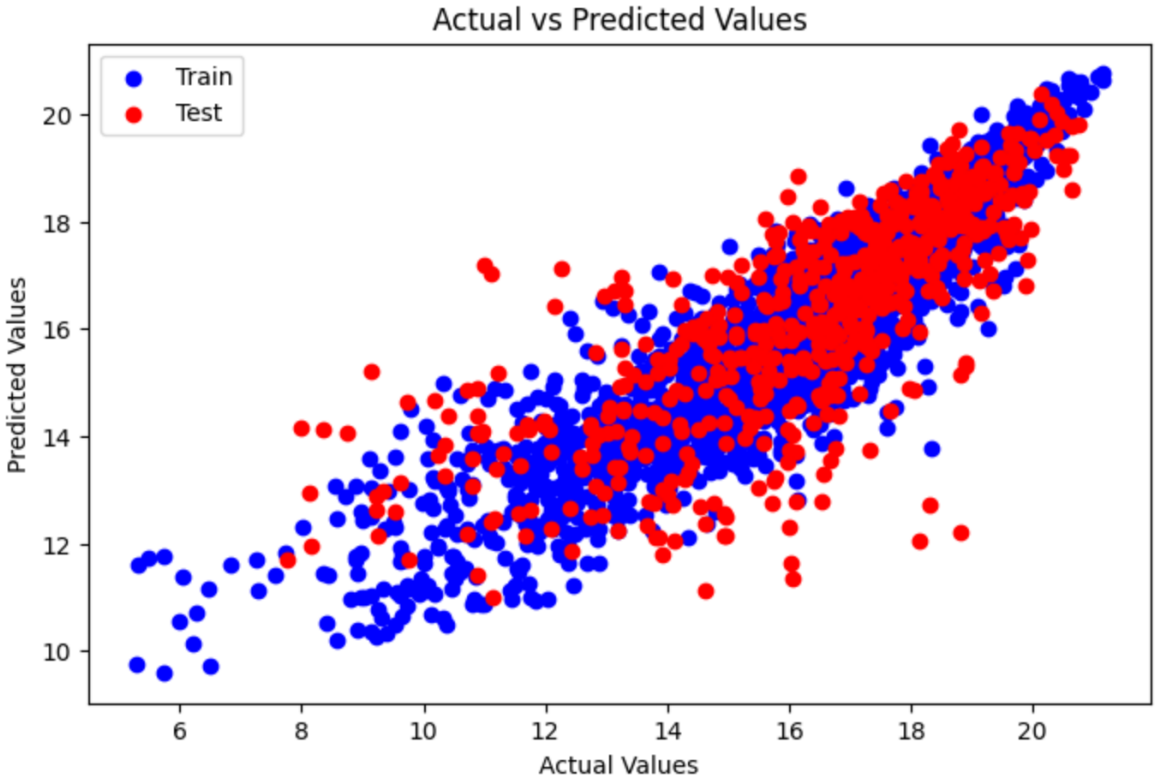
Random Forest dựa trên phương pháp ensemble learning, kết hợp nhiều mô hình học máy cơ bản để tạo ra một mô hình mạnh hơn và ổn định hơn. Random Forest đặc biệt hiệu quả trong việc giảm overfitting mà các model cây quyết định đơn lẻ thường gặp phải.

Nguyên lý hoạt động của Random Forest

- Chọn ngẫu nhiên mẫu dữ liệu: Từ tập dữ liệu gốc, chọn ngẫu nhiên nhiều mẫu dữ liệu (có thể có lặp lại) để huấn luyện từng cây quyết định. Đây được gọi là phương pháp bootstrap sampling.
- Xây dựng các cây quyết định: Mỗi cây quyết định được xây dựng từ một mẫu dữ liệu bootstrap và tại mỗi nút trong cây, chỉ một tập hợp con ngẫu nhiên của các đặc trưng (features) được xem xét để chia tách dữ liệu. Điều này giúp tăng tính đa dạng của các cây trong rừng.
- Dự đoán bằng cách bỏ phiếu hoặc trung bình: Đối với bài toán phân loại, kết quả cuối cùng được quyết định bằng cách lấy kết quả bỏ phiếu đa số từ các cây quyết định. Đối với bài toán hồi quy, kết quả được tính bằng cách lấy trung bình các dự đoán của các cây.

Sau khi chạy model, kết quả của nhóm em như sau:

```
Final Training Accuracy: 81.07%
Final Test Accuracy: 57.14%
Train MAPE: 0.06%
Test MAPE: 0.09%
Train MSE: 1.3071702355158863
Test MSE: 2.84212090939156
```




3.6 Random forests

Random Forest dựa trên phương pháp ensemble learning, kết hợp nhiều mô hình học máy cơ bản để tạo ra một mô hình mạnh hơn và ổn định hơn. Random Forest

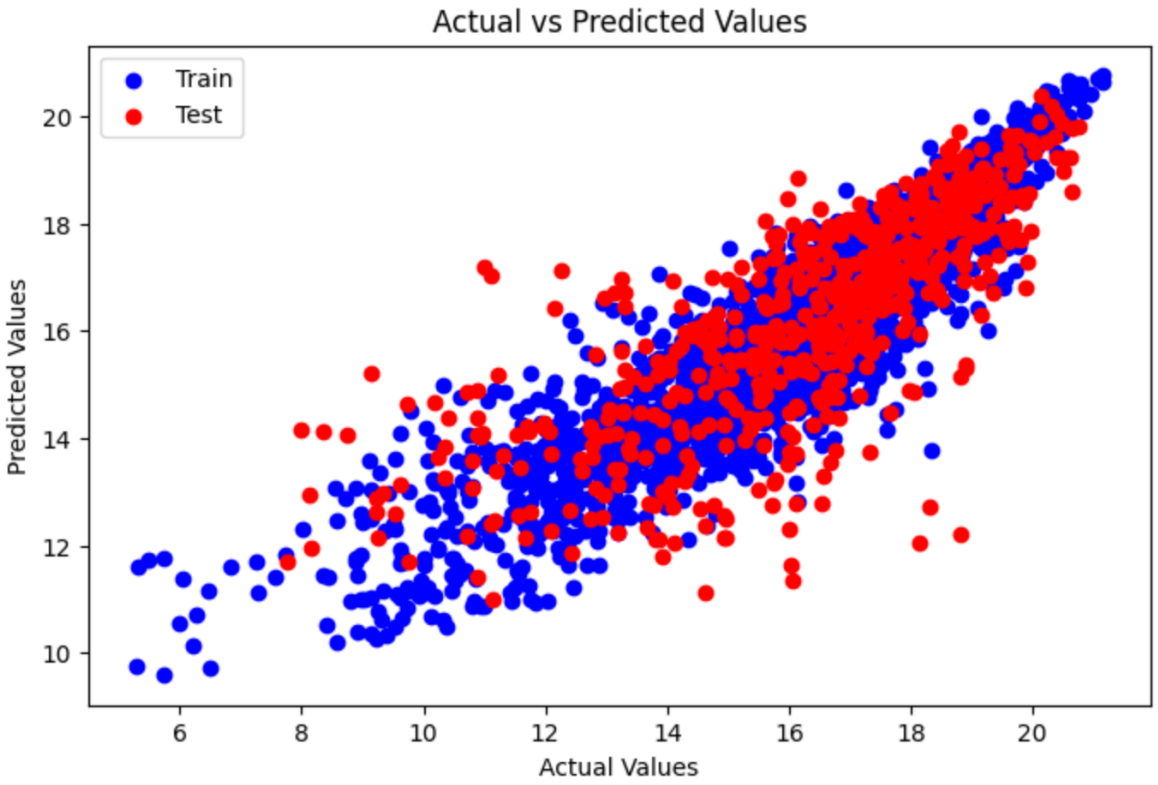
đặc biệt hiệu quả trong việc giảm overfitting mà các model cây quyết định đơn lẻ thường gặp phải.

Nguyên lý hoạt động của Random Forest

- Chọn ngẫu nhiên mẫu dữ liệu: Từ tập dữ liệu gốc, chọn ngẫu nhiên nhiều mẫu dữ liệu (có thể có lặp lại) để huấn luyện từng cây quyết định. Đây được gọi là phương pháp bootstrap sampling.
- Xây dựng các cây quyết định: Mỗi cây quyết định được xây dựng từ một mẫu dữ liệu bootstrap và tại mỗi nút trong cây, chỉ một tập hợp con ngẫu nhiên của các đặc trưng (features) được xem xét để chia tách dữ liệu. Điều này giúp tăng tính đa dạng của các cây trong rừng.
- Dự đoán bằng cách bỏ phiếu hoặc trung bình: Đối với bài toán phân loại, kết quả cuối cùng được quyết định bằng cách lấy kết quả bỏ phiếu đa số từ các cây quyết định. Đối với bài toán hồi quy, kết quả được tính bằng cách lấy trung bình các dự đoán của các cây.

Sau khi chạy model, kết quả của nhóm em như sau:

```
Final Training Accuracy: 81.07%
Final Test Accuracy: 57.14%
Train MAPE: 0.06%
Test MAPE: 0.09%
Train MSE: 1.3071702355158863
Test MSE: 2.84212090939156
```



3.7 XGBoosting

XGBoost là một phiên bản cải tiến của Gradient Boosting, tối ưu hóa cả về tốc độ và hiệu suất. Cách thức hoạt động của XGBoost

1. Xây dựng các cây quyết định

XGBoost xây dựng các cây quyết định nhỏ tuần tự. Mỗi cây quyết định học từ lỗi của các cây trước đó. Quá trình này tiếp tục cho đến khi đạt được số lượng cây quy định hoặc lỗi không giảm nữa.

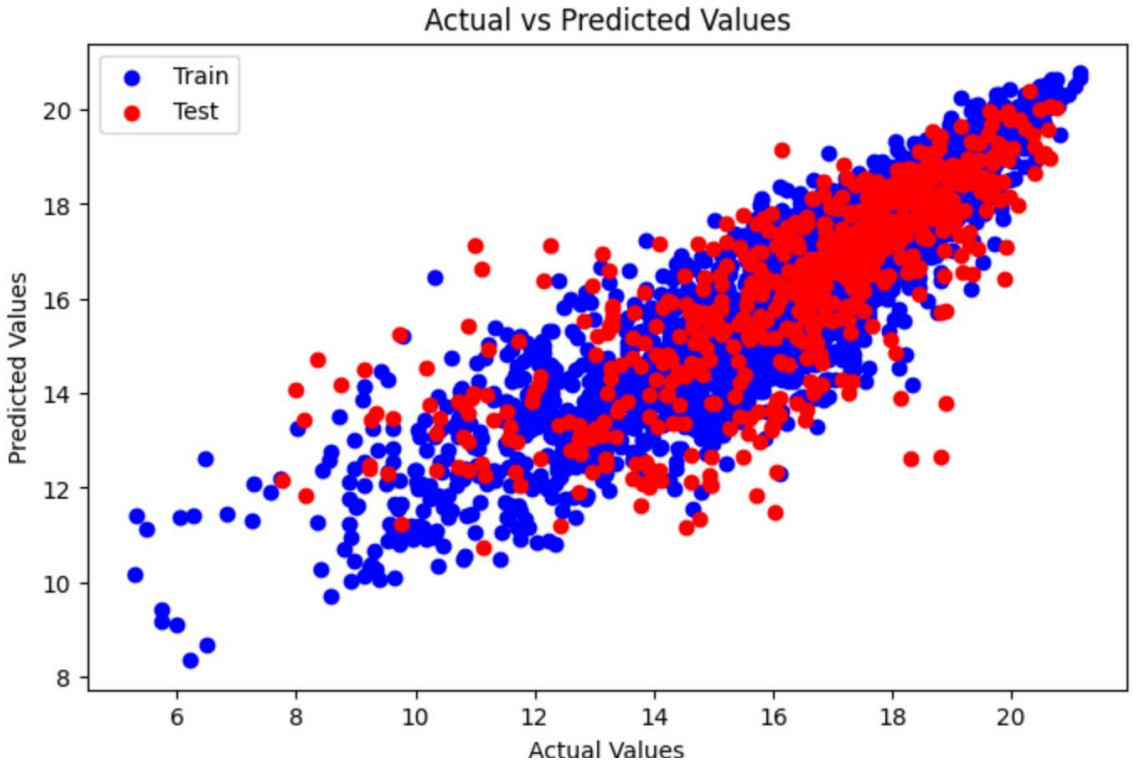
2. Tối ưu hóa

XGBoost sử dụng các kỹ thuật tối ưu hóa để giảm thiểu hàm lỗi. Nó sử dụng thuật toán gradient descent để điều chỉnh các trọng số của các mô hình một cách hiệu quả.

Sau đây là kết quả của nhóm em khi chạy model XGBoosting:

```
Best Parameters: {'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 500}
Best R^2 Score: 0.61195615532664

Final Training Accuracy: 78.08%
Final Test Accuracy: 58.44%
Train MAPE: 0.06%
Test MAPE: 0.08%
```

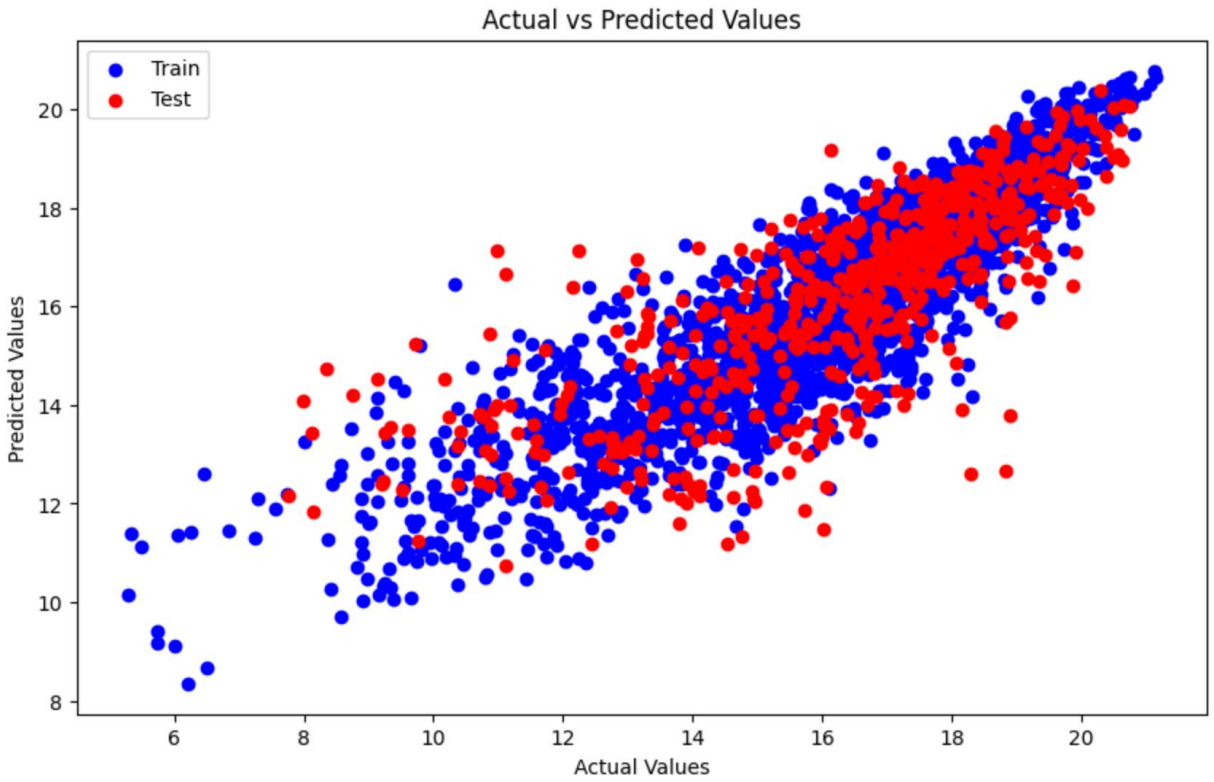


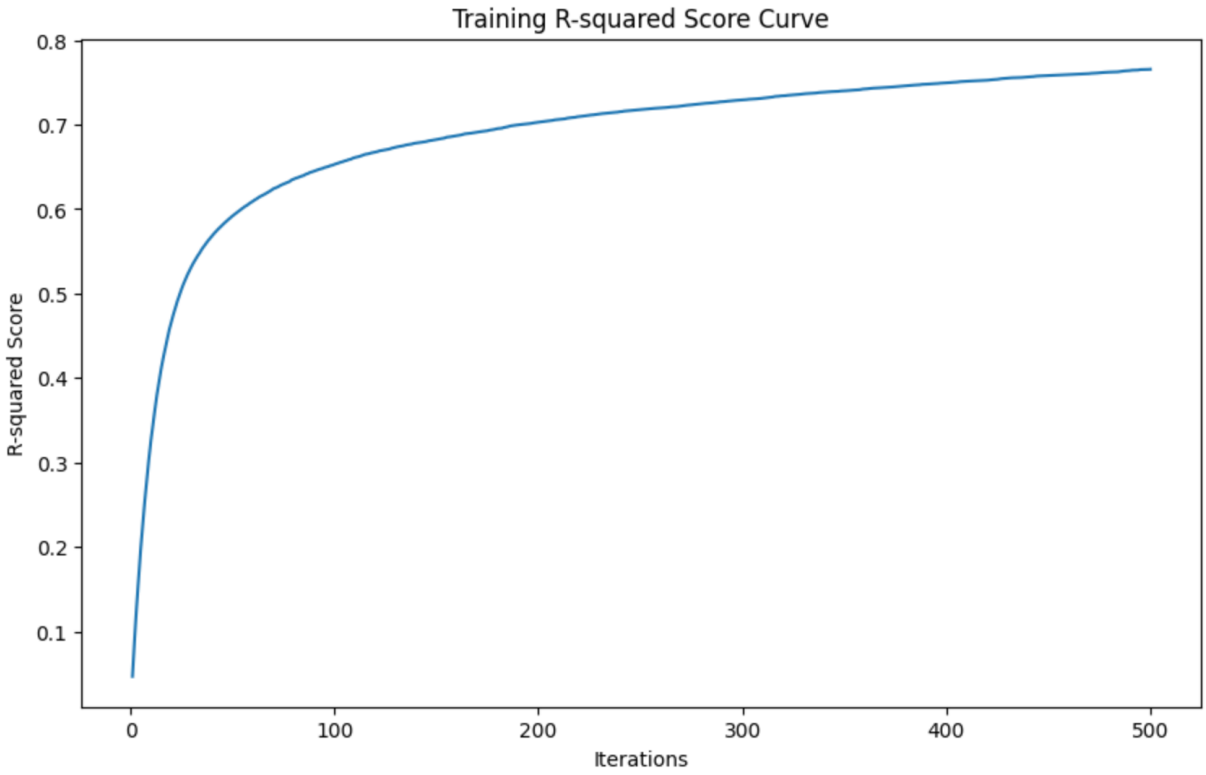
### 3.8 Tracking XGBoosting

Tracking XGBoost đề cập đến việc theo dõi và giám sát quá trình huấn luyện của mô hình XGBoost, một thuật toán mạnh mẽ dựa trên Gradient Boosting. Việc theo dõi này bao gồm việc giám sát các chỉ số hiệu suất của model, như lỗi huấn luyện, lỗi kiểm tra, và các chỉ số khác để đảm bảo rằng model đang được huấn luyện đúng cách và đạt được hiệu suất tối ưu. Kết quả của nhóm em sau khi chạy model như sau:

```
Best Parameters: {'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 500}
Best R^2 Score: 0.61195615532664

Final Training Accuracy: 78.08%
Final Test Accuracy: 58.44%
Train MAPE: 0.06%
Test MAPE: 0.08%
Train MSE: 1.5136856808393417
Test MSE: 2.7560382020452128
```






## 4. Kết luận

Mục tiêu của dự án này là dự đoán doanh thu phòng vé của một bộ phim từ dữ liệu công khai bằng cách sử dụng học máy. Kết luận có thể rút ra là các model được phát triển trong dự án này còn xa mới hoàn hảo. Nhiều biến số khác có thể đã được xem xét trong quá trình dự đoán, chẳng hạn như trượt giá, cốt truyện phim, hiệu quả của maketing, danh tiếng ngôi sao, danh tiếng đạo diễn, đề cử giải thưởng, v.v. Ngoài ra, một kỹ thuật mô hình hóa phức tạp hơn như Random forests có thể đã mang lại kết quả dự đoán doanh thu phòng vé tốt hơn.

## Phụ lục phân chia công việc

*/các phần bôi đậm trong bảng tương ứng với thành viên đảm nhiệm*

| **Thành viên** / **Công việc** | **Nguyễn Phương Trang** | **Đỗ Thị Thuỳ Trang** | **Nguyễn Minh Hường** |
|---|---|---|---|
| **Thu thập dữ liệu** | | | |
| **Tiền xử lý dữ liệu với IMDb data** | | | |

| | | | |
|---|---|---|---|
| **Tiền xử lý dữ liệu với Kaggle data** | | | |
| **Xử lý và trực quan hóa dữ liệu** | | | |
| **Xây dựng model baseline, Decision tree** | | | |
| **Cải tiến model** | | | |
| **Viết báo cáo** | | | |
| **Phần trăm đóng góp** | | | |