ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

□□□□□

PHÂN TÍCH NHỮNG YẾU TỐ ẢNH HƯỞNG ĐẾN DOANH THU PHIM

Nhóm 17			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Trương Nhật Quang	22521207	KHDL
2	Đỗ Tuấn Trực	22521548	KHDL
3	Nguyễn Đức Tài	22521277	KHDL
4	Trương Lưu Song Tâm	22521294	KHDL

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Việc phân tích những yếu tố ảnh hưởng tới doanh thu của một bộ phim là một vấn đề quan trọng, không chỉ giúp các nhà sản xuất và phát hành tối ưu hóa chiến lược kinh doanh mà còn hỗ trợ việc phân bổ ngân sách và xác định các yếu tố thành công. Trong nghiên cứu này, chúng tôi tập trung vào việc xác định các yếu tố tác động đến doanh thu của các bộ phim dựa trên các yếu tố khác nhau và đã triển khai nhiều mô hình học máy với các tiêu chí đánh giá là MSE, R^2 và MAE. Các mô hình được sử dụng bao gồm LightGBM, Random Forest, Linear Regression, Ridge Regression và XGBoost. Trong đó, **LightGBM** đạt hiệu suất tốt nhất với $R^2 = 0.299$ và sai số thấp nhất (MAE xấp xỉ 179.5 triệu USD). Kết quả phân tích tầm quan trọng của các biến cho thấy yếu tố chính ảnh hưởng đến doanh thu là **số lượt bình chọn (votes)** theo sau là **xếp hạng (rating).** Các yếu tố khác **metascore** hay các cụm từ **keywords** cũng đóng vai trò đáng kể trong việc cải thiện độ chính xác của mô hình dự đoán.

Trong nghiên cứu này, chúng tôi đã sử dụng hai bộ dữ liệu: bộ dữ liệu phân tích tự thu thập tại web [1] (đặt tên là imdb.csv, cập nhật tới năm 2023) thông qua công cụ Python BeautifulSoup, và bộ dữ liệu tham khảo có sẵn tại [2] (đặt tên là keywordsmovie.csv, cập nhật tới năm 2017). Chúng tôi cam kết minh bạch về nguồn gốc của dữ liệu sử dụng trong nghiên cứu, với bộ dữ liệu tự thu thập được ghi rõ trong tài liệu tham khảo và bộ dữ liệu tham khảo chỉ dùng làm nền tảng hỗ trợ. Bộ dữ liệu và đề tài được nhóm tự thiết kế, không dựa trên bất kỳ đề tài nào khác. Ngoài ra, nhóm cũng đã thực hiện chỉnh sửa và xây dựng thêm cột dữ liệu để phù hợp với yêu cầu nghiên cứu. Những thiếu sót, nhóm sẽ cập nhật và hoàn thiện trong tương lai.

2. MÔ TẢ BỘ DỮ LỆU

Bộ dữ liệu phân tích được thu thập tại [1]. (đặt tên là imdb.csv cập nhật tới 2023) thông qua công cụ Python BeautifulSoup.

Bộ dữ liệu phân tích được tham khảo tại [2]. (đặt tên là keywordsmovie.csv cập nhật tới 2017).

Bộ dữ liệu đầu tiên (**imdb.csv**) chứa thông tin chi tiết về 1,000 bộ phim và gồm các cột sau: "Name of movie", "Year of release", "Watchtime", "Genre", "Classification", "Movie Rating", "Metascore", "Votes", "Gross collection", "Des".

Bộ dữ liệu thứ hai (**keywordsmovie.csv**) gồm 45,465 dòng, lưu trữ các từ khóa liên quan đến phim, chỉ gồm hai cột quan trọng là: "title" (Tiêu đề của bộ phim, giống với **Name of movie** trong bộ dữ liệu imdb.csv), "Keywords".

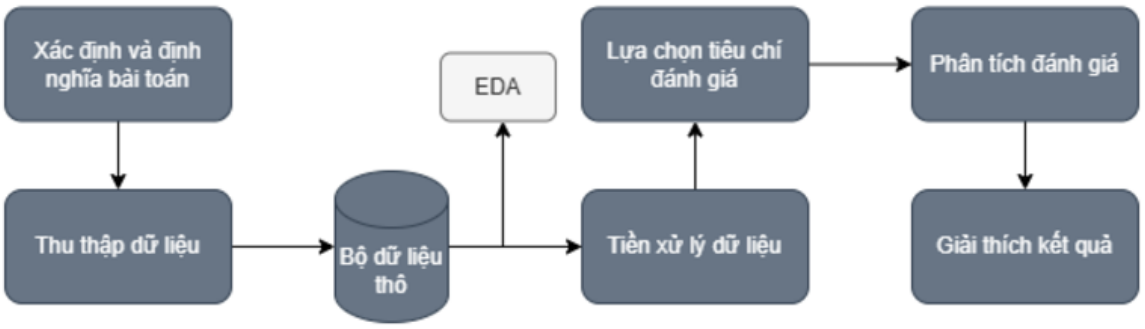
Việc có chung cột **Name of movie** trong cả hai bộ dữ liệu giúp tạo liên kết giữa chúng, đóng vai trò là khóa ngoại (foreign key). Điều này giúp tối ưu hóa hiệu suất truy vấn và cải thiện tốc độ xử lý dữ liệu, nhất là trong các ứng dụng lớn và phức tạp như hệ thống gợi ý phim. Sau đó, hai bộ dữ liệu được gộp lại thành một tập dữ liệu duy nhất có tên **movieofficial.csv**, gồm 1,000 bộ phim và 11 cột thuộc tính. Gồm 5 biến phân loại Name of movie, Genre, Classification, Des, Keywords. Và 6 biến số Year of release, Watchtime, Movie Rating, Metascore, Votes, Gross collection.

Cột dữ liệu	**Nội dung thuộc tính**	**Ví dụ**	**Kiểu dữ liệu**
Name of movie	Tiêu đề của bộ phim	Iron Man 3	Object
Year of release	Năm phát hành của bộ phim	2013	int64
Watchtime	Thời lượng của bộ phim tính bằng phút	130	int64
Genre	Thể loại của bộ phim	Action, Adventure, Sci-Fi	Object
Classification	Phân loại của bộ phim cho lứa tuổi	PG-13	Object
Movie Rating	Điểm đánh giá của bộ phim do người dùng đưa ra	7.1	float64
Metascore	Điểm metascore của bộ phim	62	float64
Votes	Số lượng phiếu bầu cho bộ phim	896,434	Int32
Gross collection	Tổng doanh thu của bộ phim	1,215,577,205	float64
Des	Mô tả nội dung của bộ phim.	When Tony Stark's world is torn apart by a formidable	Object

		terrorist called the Mandarin, he starts an odyssey of rebuilding and retribution.	
Keywords	Các từ khóa mô tả nội dung từng bộ phim	terrorist, war on terror, tennessee, malibu, marvel comic, superhero, based on comic, tony stark, iron man, aftercreditsstinger, marvel cinematic universe, mandarin, 3d, war machine, iron patriot, extremis	Object

3. PHƯƠNG PHÁP PHÂN TÍCH

3.1 Quy trình thực hiện



3.2 Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước quan trọng để làm sạch và chuẩn bị dữ liệu trước khi áp dụng các mô hình học máy. Ở bước này, dữ liệu sẽ được làm sạch và chuẩn hóa, giúp mô hình học được tốt hơn.

- **Chuẩn hóa kiểu dữ liệu**: Các giá trị trong cột **votes** và **gross** có chứa ký tự không phải số (ví dụ dấu phẩy hoặc ký tự không hợp lệ). Chúng sẽ được chuyển đổi sang kiểu dữ liệu số (int cho **votes** và float cho **gross**).
- **Xử lý giá trị thiếu (Missing Data)**: Các giá trị thiếu sẽ được xử lý bằng cách điền **giá trị trung bình** (mean) cho các cột số, hoặc **Unknown** cho các cột không phải số.

- **Mã hóa từ khóa**: Cột **keywords** được mã hóa bằng phương pháp **TF-IDF,** phương pháp này giúp chuyển mỗi từ khóa trong keywords thành một vector số, phản ánh mức độ quan trọng của mỗi từ khóa trong bộ phim so với tất cả các bộ phim khác. Nhóm sẽ chỉ sử dụng 100 từ khóa quan trọng nhất để tránh làm tăng độ phức tạp của mô hình.

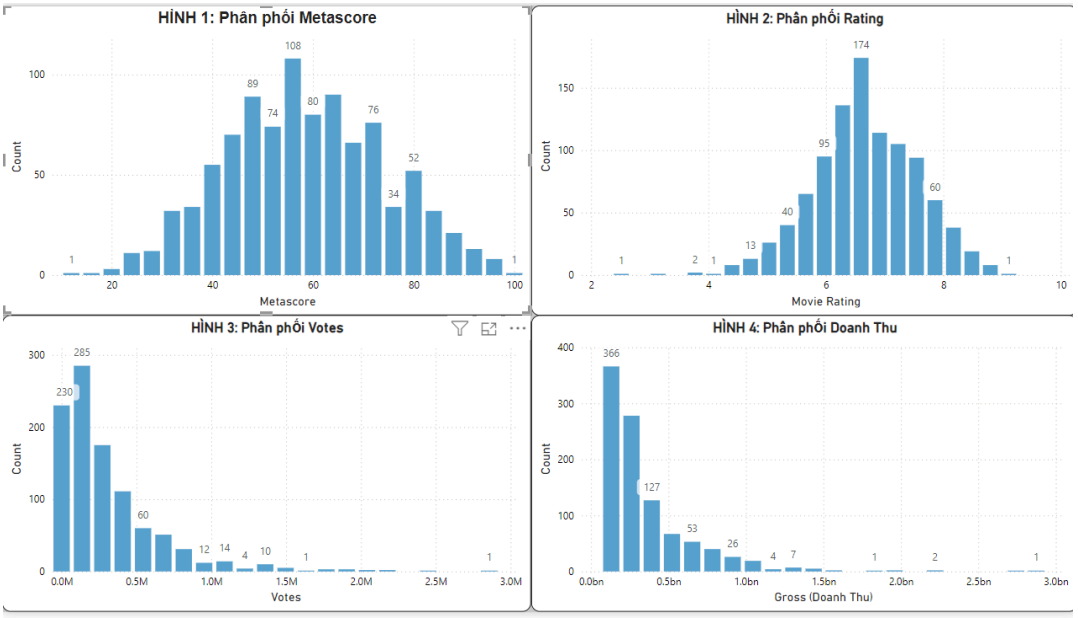
- **Giảm chiều dữ liệu**: Sử dụng **PCA** (Principal Component Analysis) để giảm số lượng đặc trưng từ các từ khóa mã hóa, giúp giảm độ phức tạp và cải thiện hiệu suất mô hình. Kỹ thuật này giúp giảm số lượng đặc trưng đầu vào mà vẫn giữ lại được phần lớn thông tin quan trọng. Ở đây, nhóm sẽ giảm chiều dữ liệu từ 100 chiều xuống 20 chiều.

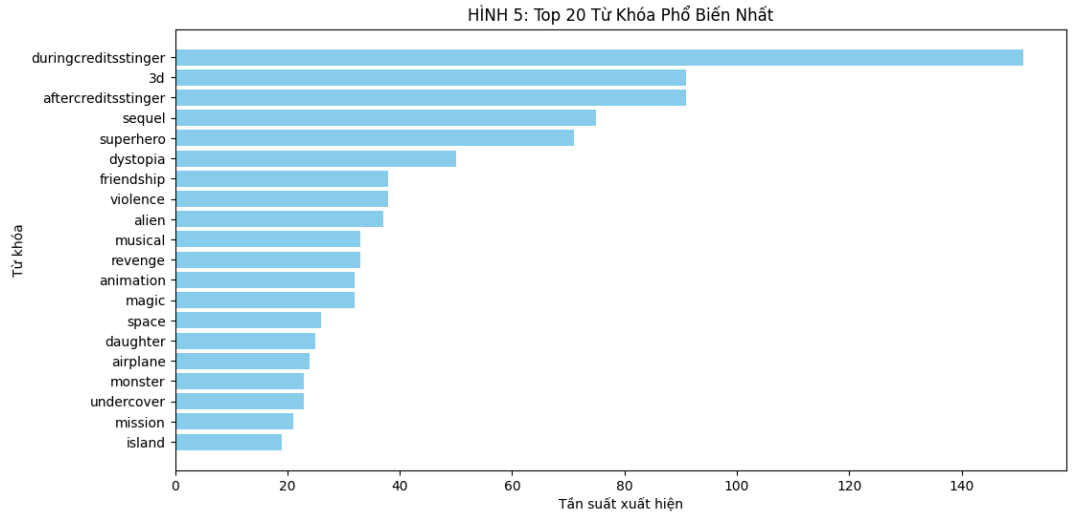
- **Chuẩn hóa dữ liệu**: Các dữ liệu sau khi được xử lý sẽ được chuẩn hóa (scaling) để đưa các giá trị về một thang đo chung, giúp mô hình học tốt hơn, tránh tình trạng các biến có giá trị lớn hơn chi phối quá mức đến mô hình.

3.3 Phân tích đặc trưng dữ liệu

Trước khi tiến hành xây dựng các mô hình đánh giá, nhóm thực hiện phân tích thăm dò dữ liệu (EDA) để hiểu rõ hơn về các mối quan hệ giữa các biến trong bộ dữ liệu. Phân tích này chủ yếu dựa trên các biểu đồ phân phối và ma trận tương quan, giúp nhận diện các yếu tố có ảnh hưởng mạnh đến doanh thu (Gross).

Biểu đồ phân phối: Các biểu đồ phân phối giúp nhận diện các xu hướng cơ bản trong dữ liệu và sự phân bổ của các biến quan trọng. Một số kết quả đáng chú ý từ các biểu đồ phân phối

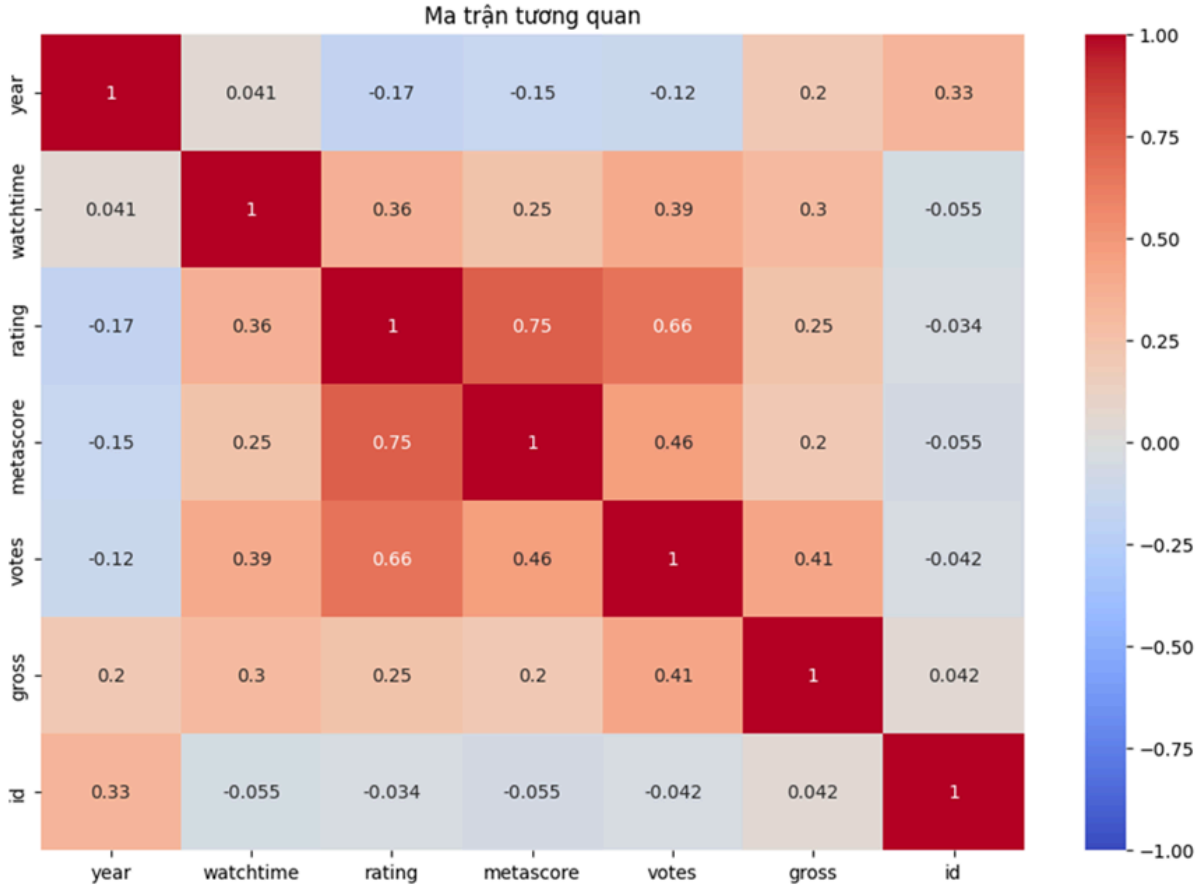




- **HÌNH 1: Phân phối Metascore** và **HÌNH 2: Phân phối rating**: Cả hai biến này đều có phân phối khá đồng đều, phân phối gần chuẩn, cho thấy rằng các bộ phim được đánh giá đa dạng cho thấy sự phân chia đồng đều trong việc đánh giá của người xem và các nhà phê bình phim.
- **HÌNH 3: Phân phối Votes**: Biểu đồ phân phối của **votes** chỉ ra có sự tập trung ở mức độ bỏ phiếu thấp, nhưng vẫn có một số bộ phim nhận được lượng bình chọn cao.
- **HÌNH 4: Phân phối doanh thu**: Biểu đồ phân phối của **gross** cho thấy lệch phải lớn, nghĩa là có một số ít bộ phim thu hút khán giả và tạo ra doanh thu cao hơn rất nhiều so với đa số bộ phim khác.
- **HÌNH 5: Top 20 từ khóa phổ biến nhất:** Các từ khóa nổi bật như duringcreditsstinger, 3D, và aftercreditsstinger xuất hiện với tần suất cao nhất, phản ánh xu hướng các bộ phim có cảnh trong và sau credits cùng công nghệ 3D thu hút sự quan tâm lớn từ khán giả. Các từ khóa còn lại

như sequel, superhero và dystopia cho thấy các bộ phim thuộc thể loại siêu anh hùng, phần tiếp theo và viễn tưởng vẫn chiếm ưu thế.

Ma trận tương quan: Ma trận tương quan cung cấp cái nhìn tổng quan về mức độ mối quan hệ giữa các biến.



- **gross** có mối quan hệ trung bình với **votes** và **rating**, với **votes** (hệ số tương quan là 0.41) và **rating** (hệ số tương quan là 0.25). Tuy mức độ không quá mạnh, nhưng lại là biến cố độ tương quan tốt nhất trong tất cả. Điều này chỉ ra rằng các bộ phim thu hút được lượng lớn khán giả bình chọn và nhận được đánh giá tích cực thường có xu hướng đạt doanh thu cao hơn. Điều này phản ánh sức hút của bộ phim không chỉ đến từ chất lượng nội dung mà còn từ sự quan tâm của khán giả.
- **Metascore** có mối tương quan yếu với gross, với hệ số tương quan là 0.2. Mặc dù mối liên hệ này không thực sự tốt, nhưng vẫn thể hiện rằng các bộ phim nhận được đánh giá tốt từ nhà phê bình có khả năng tạo ra doanh thu cao hơn ở mức độ nhất định.
- **gross** có mối quan hệ trung bình với **votes** và **rating**, nhưng nhìn chung mối quan hệ này vẫn mạnh hơn so với mối quan hệ của doanh thu đối với các biến khác, cho thấy rằng các bộ phim có sự quan tâm lớn từ người

xem và nhận được đánh giá cao từ khán giả và nhà phê bình có xu hướng có doanh thu cao hơn.

3.4 Thiết kế và đánh giá

Dựa trên các dữ liệu đã thu thập và kết quả phân tích, nhóm đã tập trung vào 4 chỉ số đánh giá bao gồm **Metascore, Votes, Keywords và Rating**. Các chỉ số này được thiết kế nhằm phản ánh mối quan hệ giữa chất lượng nội dung, mức độ quan tâm từ khán giả và các yếu tố mô tả nội dung của bộ phim đối với doanh thu. Những chỉ số này không chỉ cung cấp thông tin giá trị về tầm quan trọng của từng yếu tố mà còn giúp đưa ra các kết luận chính xác và đầy đủ hơn.

Nhóm sẽ sử dụng các mô hình học máy phổ biến để huấn luyện trên tập dữ liệu và biến mục tiêu **gross**. Các mô hình được lựa chọn bao gồm: **Random Forest Regressor, XGBoost Regressor, LightGBM Regressor, Ridge Regression Linear Regression.**

Sau khi huấn luyện, nhóm sẽ đánh giá các mô hình trên tập kiểm tra dựa trên các chỉ số hiệu suất:

- **MSE (Mean Squared Error):** Đo lường sai số bình phương trung bình giữa giá trị dự đoán và thực tế. Mô hình có MSE thấp sẽ có dự đoán gần với giá trị thực tế.
- **R^2 Score:** Đánh giá khả năng giải thích sự biến động của dữ liệu. Giá trị R^2 càng gần 1, mô hình càng tốt.
- **MAE (Mean Absolute Error):** Tính sai số tuyệt đối trung bình, thể hiện độ chính xác của mô hình. Mô hình có MAE thấp sẽ có sai số tuyệt đối trung bình thấp, nghĩa là dự đoán chính xác.

Đánh giá tầm quan trọng của các biến: Sau khi so sánh hiệu suất và tìm được mô hình tốt nhất, nhóm sẽ sử dụng mô hình đó để tính toán và đánh giá tầm quan trọng của các đặc trưng. Sau đó trực quan hóa kết quả vừa thu được, từ đó dễ dàng nhận diện được những biến ảnh hưởng nhiều nhất đến dự đoán doanh thu.

4. KẾT QUẢ PHÂN TÍCH

4.1 Kết quả phân tích từ các mô hình

Nhằm kiểm tra và đánh giá khả năng đánh giá những yếu tố ảnh hưởng tới doanh thu (Gross), nhóm đã tiến hành huấn luyện các mô hình học máy bao gồm Random Forest Regressor, XGBoost Regressor, LightGBM Regressor, Ridge Regression và Linear Regression.

Kết quả thu được từ quá trình huấn luyện và đánh giá mô hình:

Model	R^2	MSE	MAE
Linear Regression	**0.204667**	**7.298993e+16**	**1.907703e+08**
Ridge Regression	**0.204911**	**7.296752e+16**	**1.907300e+08**
Random Forest	**0.230611**	**7.060897e+16**	**1.767485e+08**
XGBoost	**0.141388**	**7.879717e+16**	**1.892369e+08**
LightGBM	**0.299325**	**6.430285e+16**	**1.794807e+08**

Trong các mô hình trên nhóm nhận thấy:

- **LightGBM:** hiệu suất tốt nhất với $R^2 = 0.299$, sai số thấp nhất (MAE xấp xỉ 179.5 triệu USD), phù hợp nhất cho dự đoán doanh thu.
- **Ridge Regression:** kết quả ổn định với $R^2 = 0.205$, sai số trung bình (MAE xấp xỉ 190.7 triệu USD).
- **Linear Regression:** hiệu suất trung bình với $R^2 = 0.205$, sai số (MAE xấp xỉ 190.8 triệu USD), tương đương Ridge Regression.
- **Random Forest:** kết quả kém hơn với $R^2 = 0.231$, sai số trung bình (MAE xấp xỉ 176.7 triệu USD), dự đoán chưa thực sự tốt.
- **XGBoost:** hiệu suất thấp nhất với $R^2 = 0.141$, sai số lớn nhất (MAE xấp xỉ 189.2 triệu USD), dự đoán chưa thực sự tốt.

Kết luận: Mặc dù độ sai số vẫn cao kể cả năm mô hình nhưng nhóm vẫn quyết định chọn ra mô hình tốt nhất là **LightGBM** để **Phân tích tầm quan trọng.**

4.2 Phân tích tầm quan trọng

Chỉ số tầm quan trọng (importance) của các biến là thước đo phản ánh mức độ ảnh hưởng của mỗi biến (hoặc đặc trưng) trong việc dự đoán biến mục tiêu.

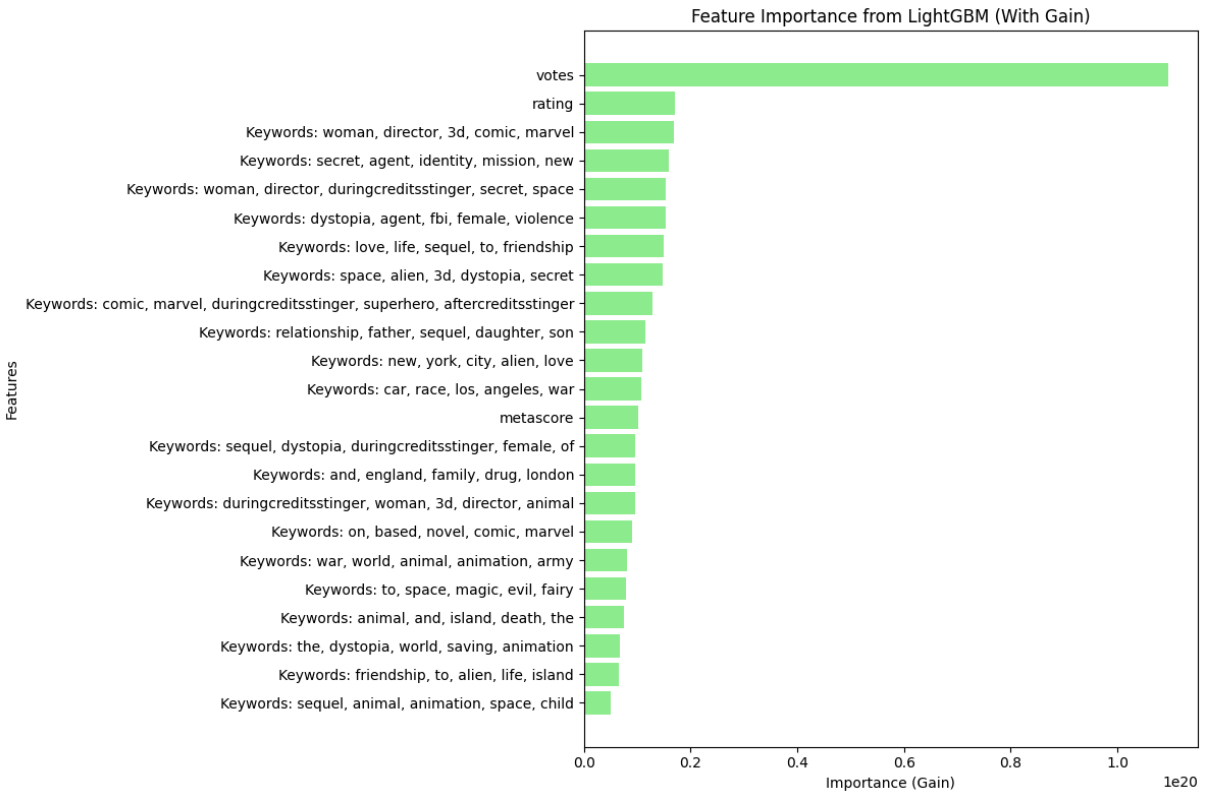
Các loại tầm quan trọng:

- **Weight** đo lường tần suất mà một đặc trưng được sử dụng trong các phân chia cây quyết định.
- **Gain** đo lượng sự cải thiện trong hàm mất mát khi đặc trưng đó được sử dụng.
- **Cover** đo lường sự tác động của đặc trưng đến số lượng mẫu (data points) được phân chia.

Trong đó, **Gain** được coi là chỉ số quan trọng nhất vì nó đo lường ảnh hưởng của mỗi một đặc trưng đóng trong việc giảm độ lỗi của mô hình. Vì vậy trong nguyên

cứu này, nhóm xác định sử dụng Gain làm chỉ số chính, vì nó phản ánh mức độ quan trọng của mỗi đặc trưng vào việc cải thiện hiệu suất mô hình.

Bảng tầm quan trọng của các biến từ mô hình **LightGBM**:



Từ bảng nhóm phân tích:

- **Votes**: Đây là đặc trưng quan trọng nhất trong việc dự đoán doanh thu phim, với mức ảnh hưởng vượt trội. Điều này khẳng định rằng số lượng bình chọn từ khán giả phản ánh mức độ quan tâm và sự phổ biến của bộ phim, qua đó tác động trực tiếp đến doanh thu.
- **Rating**: Đứng thứ hai về mức độ quan trọng, rating cho thấy đánh giá của khán giả có vai trò then chốt trong việc dự đoán doanh thu. Đây là thước đo quan trọng về chất lượng cảm nhận của khán giả đối với phim.
- **Metascore**: Mặc dù không chiếm vị trí cao như votes và rating, metascore cũng góp phần đáng kể, thể hiện sự ảnh hưởng của đánh giá từ các nhà phê bình đối với doanh thu.
- **Keywords**: Nhiều từ khóa liên quan đến nội dung, thể loại phim đóng vai trò quan trọng trong việc dự đoán. Các từ khóa nổi bật bao gồm:
- **"woman, director, 3D, comic, marvel"**: Thể hiện sức hút mạnh mẽ từ các bộ phim thuộc vũ trụ điện ảnh Marvel.
- **"secret, agent, identity, mission, new"**: Tập trung vào thể loại hành động, gián điệp và bí ẩn, thu hút sự quan tâm lớn từ khán giả.

- **"dystopia, agent, FBI, female, violence"**: Thể hiện sự phổ biến của các bộ phim với chủ đề dystopia, bạo lực và nhân vật nữ chính mạnh mẽ.
- **"space, alien, 3D, dystopia, secret"**: Nhấn mạnh sức hút của các bộ phim khoa học viễn tưởng và bí ẩn.
- **"comic, marvel, duringcreditsstinger, superhero"**: Xác nhận vị trí nổi bật của các bộ phim siêu anh hùng trong việc tạo ra doanh thu lớn.

Nhận xét: có thể kết hợp các yếu tố lượng bình chọn, đánh giá chất lượng và từ khóa mô tả nội dung từ đó tối ưu cách phát triển phim để phù hợp hơn với thị trường kiếm thêm doanh thu.

5. KẾT LUẬN

Nhóm đã tiến hành thu thập, tiền xử lý dữ liệu và xây dựng các mô hình để phân tích các yếu tố ảnh hưởng tới doanh thu của bộ phim doanh thu (Gross) của bộ phim. Sau khi huấn luyện và đánh giá các mô hình bao gồm Random Forest, XGBoost, LightGBM, Ridge Regression và Linear Regression, kết quả cho thấy LightGBM có hiệu suất tốt nhất với $\mathbf{R^2 = 0.299}$ và sai số thấp nhất (MAE xấp xỉ **179.5 triệu USD**), khẳng định khả năng phân tích vượt trội trong việc xác định các yếu tố quan trọng so với các mô hình còn lại.

Phân tích tầm quan trọng của các đặc trưng từ mô hình LightGBM cho thấy **Votes** là yếu tố quan trọng nhất, thể hiện rằng lượng bình chọn từ khán giả đóng vai trò lớn nhất trong việc ảnh hưởng đến doanh thu. Tiếp theo là **Rating**, phản ánh tầm quan trọng của đánh giá từ khán giả và nhà phê bình. Các từ khóa (**Keywords**) liên quan đến nội dung như **"woman, director, 3D, comic, marvel"**, **"secret, agent, identity, mission, new"**, và **"space, alien, 3D, dystopia, secret"** và **Metascore** cũng có mức độ ảnh hưởng đáng kể. Những từ khóa này thể hiện xu hướng ưa chuộng của khán giả đối với các thể loại phim như khoa học viễn tưởng, siêu anh hùng, gián điệp, và hành động. Từ kết quả này, nhóm nhận thấy để tối ưu doanh thu, nhà sản xuất phim cần tập trung vào các chiến lược thu hút sự quan tâm và bình chọn từ khán giả, nâng cao chất lượng nội dung thuộc các thể loại được yêu thích, và tận dụng các yếu tố đặc biệt như 3D hoặc các phân đoạn after-credits để tạo sức hút. Đồng thời, việc đầu tư vào các yếu tố sáng tạo và chiến lược quảng bá nhằm tăng đánh giá từ cả khán giả lẫn nhà phê bình cũng là yếu tố then chốt.

Tuy nhiên, nhóm nhận thấy các mô hình vẫn chưa đạt được độ chính xác như kỳ vọng. Sai số dự đoán còn khá cao, dẫn đến độ lệch doanh thu lớn trong một số trường hợp. Một thách thức lớn khác là việc xử lý và mã hóa các từ khóa (**Keywords**) để đưa vào mô hình học máy. Do dữ liệu từ khóa mang tính chất không có cấu trúc, việc trích xuất và đánh giá mối quan hệ giữa từ khóa với doanh thu còn gặp nhiều khó khăn. Các kết quả thu được mang tính tương đối, chưa thể phản ánh đầy đủ mối quan

hệ thực sự giữa các từ khóa và biến mục tiêu. Nhóm đề xuất các hướng nghiên cứu tương lai, bao gồm cải thiện kỹ thuật xử lý văn bản và áp dụng các mô hình ngôn ngữ tiên tiến để nâng cao độ chính xác trong phân tích.

TÀI LIỆU THAM KHẢO

[1] IMDb : https://www.imdb.com/list/ls098063263/

[2] Kaggle : Getting Started with a Movie Recommendation System
https://www.kaggle.com/code/ibtesama/getting-started-with-a-movie-recommendation-system/input?select=keywords.csv

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Trương Nhật Quang	SLIDE, Viết giới thiệu, mô tả, phân tích và kết quả, kết luận, thuyết trình
2	Nguyễn Đức Tài	Viết phân tích và kết quả, kết luận, thuyết trình
3	Đỗ Tuấn Trực	Viết phân tích và kết quả, kết luận, thuyết trình
4	Trương Lưu Song Tâm	Viết giới thiệu, mô tả, thuyết trình