**TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**

**KHOA CÔNG NGHỆ THÔNG TIN**

---------\*\*\*\*\*---------

A blue circle with white text and a globe

Description automatically generated

**BÁO CÁO**

**BỘ MÔN ĐỒ ÁN 1**

***Đề tài:***

**DỰ ĐOÁN ĐỘ THÀNH CÔNG CỦA PHIM   
CHIẾU RẠP TẠI VIỆT NAM**

|  |  |
| --- | --- |
| **GIẢNG VIÊN HƯỚNG DẪN**  **Nguyễn Thị Mai Dung** | **SINH VIÊN THỰC HIỆN**  **Khổng Thị Hoà Phan Văn Huy Đinh Ngọc Khuê** |

***Hà Nội - 2025***

TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

KHOA CÔNG NGHỆ THÔNG TIN

**---------\*\*\*\*\*---------**

**A blue circle with white text and a globe

Description automatically generated**

**BÁO CÁO**

**BỘ MÔN ĐỒ ÁN 1**

|  |  |
| --- | --- |
| Sinh viên thực hiện : Khổng Thị Hoà | MSSV: 2321050043 |
| Sinh viên thực hiện : Phan Văn Huy | MSSV: 2321050069 |
| Sinh viên thực hiện : Đinh Ngọc Khuê | MSSV: 2321050065 |

***Đề tài :***

**DỰ ĐOÁN ĐỘ THÀNH CÔNG CỦA PHIM   
CHIẾU RẠP TẠI VIỆT NAM**

|  |
| --- |
| GIẢNG VIÊN HƯỚNG DẪN : Nguyễn Thị Mai Dung |

***Hà Nội – 2025***

MỤC LỤC

LỜI MỞ ĐẦU

[1 PHẦN 1: GIỚI THIỆU CHUNG 2](#_Toc209573642)

[1.1 Tổng quan về bài toán dự đoán trong ngành điện ảnh 2](#_Toc209573643)

[1.2 Các yếu tố ảnh hưởng đến thành công của phim 2](#_Toc209573644)

[1.3 Ý nghĩa khoa học và thực tiễn của đề tài 3](#_Toc209573645)

[2 PHẦN 2: CƠ SỞ LÝ THUYẾT 4](#_Toc209573646)

[2.1 Cơ sở căn bản 4](#_Toc209573647)

[2.2 Khái niệm cơ bản về Machine Learning và bài toán Phân loại 5](#_Toc209573648)

[2.3 Các chỉ số đánh giá hiệu quả mô hình 5](#_Toc209573649)

[2.4 Các thuật toán dự đoán được sử dụng 7](#_Toc209573650)

[3 PHẦN 3: PHÂN TÍCH DỮ LIỆU VÀ THIẾT KẾ MÔ HÌNH 8](#_Toc209573651)

[3.1 Giới thiệu về dữ liệu 8](#_Toc209573652)

[3.2 Tiền xử lý dữ liệu 8](#_Toc209573653)

[3.3 Tạo nhãn thành công 9](#_Toc209573654)

[3.4 Phân tích dữ liệu khám phá (EDA) 10](#_Toc209573655)

[3.5 Chuẩn bị dữ liệu cho mô hình 11](#_Toc209573656)

[3.6 Thiết kế mô hình 11](#_Toc209573657)

[4 PHẦN 4: PHƯƠNG PHÁP THỰC HIỆN VÀ KẾ HOẠCH ĐÁNH GIÁ 12](#_Toc209573658)

[4.1 Kế hoạch triển khai mô hình 12](#_Toc209573659)

[4.2 Phương pháp đánh giá 12](#_Toc209573660)

[4.3 Phương pháp kiểm chứng mô hình dự kiến 13](#_Toc209573661)

[4.4 Feature Importance Analysis 13](#_Toc209573662)

[4.5 Kế hoạch phân tích lỗi 14](#_Toc209573663)

[4.6 Thách thức dự kiến và giải pháp 14](#_Toc209573664)

[5 PHẦN 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 16](#_Toc209573665)

[5.1 Tổng kết 16](#_Toc209573666)

[5.2 Ý nghĩa và tác động thực tiễn 16](#_Toc209573667)

[5.3 Nhận thức về hạn chế 16](#_Toc209573668)

[5.4 Lời kết 16](#_Toc209573669)

LỜI MỞ ĐẦU

Ngành công nghiệp điện ảnh là một lĩnh vực đầy tính nghệ thuật nhưng cũng chịu sự chi phối mạnh mẽ của các yếu tố kinh tế. Hàng năm, hàng trăm bộ phim được sản xuất và ra mắt công chúng, nhưng không phải dự án nào cũng đạt được thành công về mặt doanh thu hay được giới phê bình đánh giá cao. Việc dự đoán khả năng thành công của một bộ phim trước khi ra mắt là một bài toán phức tạp nhưng lại có ý nghĩa thực tiễn to lớn. Khả năng này không chỉ giúp các nhà sản xuất điện ảnh đưa ra quyết định đầu tư sáng suốt mà còn hỗ trợ các nhà phân phối, quảng bá trong việc xây dựng chiến lược tiếp thị hiệu quả.

Trong bối cảnh khoa học dữ liệu và học máy ngày càng phát triển, các phương pháp dự đoán định lượng dựa trên dữ liệu lịch sử đã trở nên khả thi hơn bao giờ hết. Thay vì chỉ dựa vào kinh nghiệm hay trực giác, các mô hình dự đoán có thể phân tích hàng loạt yếu tố như kinh phí sản xuất (Budget), thể loại (Genre), quốc gia sản xuất (Country), danh tiếng của đạo diễn và diễn viên, để đưa ra những dự báo có căn cứ. Mặc dù đã có nhiều nghiên cứu về vấn đề này, việc xây dựng một mô hình tối ưu, phù hợp với đặc thù dữ liệu điện ảnh và có khả năng giải thích tốt vẫn là một thách thức.

Đồ án này tập trung giải quyết bài toán dự đoán độ thành công của phim chiếu rạp. Mục tiêu chính là xây dựng một mô hình học máy hiệu quả để phân loại một bộ phim có thành công hay không. Để đạt được mục tiêu đó, chúng em sẽ tiến hành thu thập và xử lý dữ liệu thống kê những bộ phim đã chiếu tại Việt Nam và phân tích. Đồ án cũng sẽ tập trung vào việc xác định các yếu tố quan trọng nhất ảnh hưởng đến thành công của phim, từ đó cung cấp những cái nhìn sâu sắc cho người làm trong ngành. Hơn nữa, chúng em sẽ tiến hành so sánh và đánh giá các thuật toán học máy khác nhau để tìm ra mô hình dự đoán tối ưu nhất. Do hạn chế về thời gian và kinh nghiệm, bài báo cáo khó tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý của quý thầy cô để báo cáo được hoàn thiện hơn.

Em xin trân trọng cảm ơn!

1. PHẦN 1: GIỚI THIỆU CHUNG
   1. Tổng quan về bài toán dự đoán trong ngành điện ảnh

Ngành công nghiệp điện ảnh là một thị trường đầy tính nghệ thuật và rủi ro. Với chi phí sản xuất và quảng bá ngày càng tăng, các nhà làm phim và nhà đầu tư luôn tìm kiếm giải pháp để tối ưu hóa quyết định của mình. Thay vì chỉ dựa vào kinh nghiệm hay trực giác, việc sử dụng dữ liệu để dự đoán khả năng thành công của một bộ phim đã trở thành một hướng đi mới đầy tiềm năng. Đề tài này giải quyết bài toán đó bằng cách tiếp cận từ góc độ Khoa học Dữ liệu, biến việc dự đoán thành một bài toán phân loại (classification problem), nhằm xác định một bộ phim có khả năng thành công hay không dựa trên các đặc điểm định lượng.

Với tư cách là một sinh viên Khoa học Dữ liệu, việc thực hiện đề tài này không chỉ giúp chúng em làm quen với một bài toán thực tế mà còn là cơ hội để vận dụng các kiến thức chuyên ngành. Bài toán này đòi hỏi sự kết hợp chặt chẽ giữa việc thu thập, tiền xử lý dữ liệu (data cleaning), phân tích khám phá (EDA - Exploratory Data Analysis), và xây dựng các mô hình dự đoán. Đây là những kỹ năng cốt lõi mà một nhà khoa học dữ liệu cần có để giải quyết các vấn đề kinh doanh phức tạp.

* 1. Các yếu tố ảnh hưởng đến thành công của phim

Thành công của một bộ phim không phải là ngẫu nhiên mà là sự tổng hòa của nhiều yếu tố. Từ góc độ phân tích dữ liệu, các yếu tố này có thể được xem là các biến đầu vào (features) cho mô hình dự đoán. Chúng bao gồm:

* Yếu tố nội tại (Internal Features): Liên quan đến các đặc điểm cốt lõi của bộ phim như kinh phí sản xuất (Budget), thể loại phim, thời lượng. Các yếu tố này thường được sử dụng để xây dựng các mô hình dự đoán cơ bản
* Yếu tố nhân sự (Human Capital): Sức ảnh hưởng của đạo diễn, biên kịch, và diễn viên chính. Trong bối cảnh phân tích dữ liệu, các yếu tố này có thể được mã hóa bằng các phương pháp như one-hot encoding hoặc được xử lý bằng cách tạo ra các biến mới như số lượng giải thưởng mà họ đã đạt được
* Yếu tố thị trường (Market Factors): Bao gồm thời điểm ra mắt, sự cạnh tranh với các bộ phim khác cùng thời điểm, và quốc gia sản xuất. Phân tích các yếu tố này giúp chúng ta hiểu rõ hơn về bối cảnh kinh tế và xã hội của thị trường điện ảnh

Việc phân tích chuyên sâu từng yếu tố này không chỉ giúp xây dựng mô hình dự đoán chính xác mà còn cung cấp những hiểu biết quan trọng về các xu hướng và động lực thúc đẩy thành công trong ngành.

* 1. Ý nghĩa khoa học và thực tiễn của đề tài

**Về mặt khoa học:** Đề tài là một minh chứng cho thấy cách một bài toán kinh tế có thể được giải quyết bằng các phương pháp khoa học dữ liệu. Nó cho phép em thực hành toàn bộ quy trình của một dự án Machine Learning, từ lúc xác định vấn đề, thu thập dữ liệu, đến lựa chọn và tối ưu hóa mô hình. Đồ án sẽ sử dụng và so sánh các mô hình học máy như Hồi quy Logistic (Logistic Regression), Rừng ngẫu nhiên (Random Forest) từ đó đánh giá ưu nhược điểm của từng thuật toán trong bối cảnh cụ thể của ngành điện ảnh.

**Về mặt thực tiễn:** Mô hình dự đoán có thể trở thành một công cụ hữu ích cho các bên liên quan trong ngành công nghiệp điện ảnh, giúp họ đưa ra các quyết định dựa trên dữ liệu. Ví dụ, các nhà đầu tư có thể sử dụng mô hình để đánh giá rủi ro, trong khi các nhà tiếp thị có thể sử dụng kết quả phân tích để nhắm đúng đối tượng khán giả.

1. PHẦN 2: CƠ SỞ LÝ THUYẾT
   1. Cơ sở căn bản

Để giải bài toán này, chúng em đã bắt đầu từ việc xây dựng một nền móng lý thuyết vững chắc. Đây không chỉ là bước đi học thuật mà còn là việc trang bị một bản đồ tư duy chi tiết, giúp chúng em định hình con đường đi từ dữ liệu thô đến mô hình dự đoán. Chương này sẽ là kim chỉ nam, hệ thống hóa các khái niệm cốt lõi, công cụ đánh giá, và các thuật toán then chốt, tạo tiền đề vững chắc cho toàn bộ quá trình thực nghiệm ở các chương sau.

* + 1. A diagram of a diagram

       AI-generated content may be incorrect.Các bước thực hiện
  1. Khái niệm cơ bản về Machine Learning và bài toán Phân loại

Đối với ngành Khoa học dữ liệu, chắc hẳn chúng ta đều đọc hoặc nghe qua về Machine Learning. Học máy (Machine Learning) nổi lên như một phương pháp đột phá, nó là một chủ đề được nhắc đến rất nhiều trong thời gian trở lại đây bên cạnh trí tuệ nhân tạo, nó được ứng dụng cực kỳ nhiều ở thời điểm hiện tại trong hầu hết tất cả các lĩnh vực không chỉ riêng Khoa học dữ liệu. Machine Learning cho phép máy tính tự học hỏi từ những dữ liệu lịch sử để tìm ra các quy luật tiềm ẩn. Thay vì dựa vào những quy tắc cứng nhắc mà mình đặt cố định ra, mô hình sẽ tự mình "nhận diện" các mẫu hình và đưa ra dự đoán. Đối với đề tài “dự đoán độ thành công của phim” này, mô hình sẽ được huấn luyện trên hàng trăm bộ phim đã phát hành có đầy đủ thông tin cần thiết, từ đó học được mối quan hệ giữa các yếu tố đầu vào (ngân sách, thể loại, diễn viên,..) và kết quả đầu ra (thành công hay thất bại).

Bài toán dự đoán độ thành công của phim, với mục tiêu phân loại một bộ phim là "thành công" hoặc "không thành công", chính là một ví dụ điển hình của bài toán Phân loại (Classification). Đây là một loại hình học có giám sát (Supervised Learning) mà trong đó mô hình được huấn luyện để gán một điểm dữ liệu vào một trong các nhãn đã biết trước. Cụ thể, chúng em sẽ giải quyết bài toán phân loại nhị phân (binary classification) với hai nhãn rời rạc, biến một vấn đề phức tạp thành một bài toán toán học có thể giải quyết được:

* Lớp 1: Phim thành công
* Lớp 0: Phim không thành công
  1. Các chỉ số đánh giá hiệu quả mô hình

Việc xây dựng một mô hình chỉ là bước khởi đầu. Để thực sự hiểu được sức mạnh của nó, chúng ta cần một bộ công cụ đánh giá khách quan. Các chỉ số này không chỉ đo lường độ chính xác tổng thể mà còn đi sâu vào phân tích các loại lỗi, giúp chúng em lựa chọn mô hình tối ưu nhất cho bài toán.

* + 1. Ma trận nhầm lẫn (Confusion Matrix)

Đây là nền tảng để tính toán tất cả các chỉ số còn lại. Ma trận nhầm lẫn cung cấp một cái nhìn chi tiết về hiệu suất của mô hình bằng cách phân loại kết quả dự đoán thành bốn trường hợp:

* True Positive (TP): Mô hình dự đoán đúng một bộ phim thành công
* True Negative (TN): Mô hình dự đoán đúng một bộ phim thất bại
* False Positive (FP): Mô hình dự đoán thành công nhưng phim thực tế thất bại. Đây là "lỗi đầu tư", mang rủi ro tài chính cao
* False Negative (FN): Mô hình dự đoán thất bại nhưng phim thực tế lại thành công. Đây là "lỗi bỏ lỡ cơ hội", một tổn thất không hề nhỏ
  + 1. Các chỉ số đánh giá chính

Từ ma trận nhầm lẫn (Confusion Matrix), chúng em sẽ tính toán các chỉ số sau: Accuracy (Độ chính xác tổng thể), Precision (Độ chính xác dự báo dương tính), Recall (Độ bao phủ), F1-score. Mỗi chỉ số là một yếu tố quan trọng để quyết định mô hình có hoạt động hiệu quả hay không.

* **Accuracy (Độ chính xác tổng thể):** Tỷ lệ dự đoán đúng trên toàn bộ tập dữ liệu. Mặc dù dễ hiểu, chỉ số này có thể gây hiểu lầm nếu dữ liệu mất cân bằng
* **Precision (Độ chính xác dự báo dương tính):** Trong số những bộ phim được mô hình dự đoán là thành công, có bao nhiêu phim thực sự thành công. Chỉ số này đặc biệt quan trọng với các nhà đầu tư, giúp họ tối thiểu hóa rủi ro đầu tư sai lầm
* **Recall (Độ bao phủ):** Trong số những bộ phim thực sự thành công, có bao nhiêu bộ phim được mô hình dự đoán đúng. Chỉ số này quan trọng khi chúng ta không muốn bỏ sót bất kỳ một bộ phim tiềm năng nào.

* **F1-score**: Là trung bình điều hòa của Precision và Recall. Chỉ số này cung cấp một cái nhìn cân bằng về hiệu suất mô hình, đặc biệt hữu ích khi chúng ta không muốn thiên về tối ưu hóa một trong hai chỉ số trên
  1. Các thuật toán dự đoán được sử dụng

Để giải quyết bài toán phân loại nhị phân này, chúng em đã lựa chọn và so sánh những thuật toán học máy phổ biến, đại diện cho ba phương pháp tiếp cận khác nhau:

* + 1. Hồi quy Logistic (Logistic Regression)

Đây là mô hình thống kê tuyến tính và là lựa chọn đầu tiên để thiết lập đường cơ sở (baseline) cho đề tài. Mặc dù đơn giản, thuật toán này rất hiệu quả trong việc xác định mối quan hệ tuyến tính giữa các đặc trưng (như ngân sách, chi phí marketing) và khả năng thành công của phim, giúp chúng em có một thước đo ban đầu để so sánh với các thuật toán phức tạp hơn.

* + 1. Rừng ngẫu nhiên (Random Forest)

Để giải quyết những hạn chế của mô hình tuyến tính, chúng em đã lựa chọn Random Forest, một thuật toán học tổ hợp mạnh mẽ. Bằng cách xây dựng một "rừng" gồm nhiều cây quyết định và sử dụng kết quả "bỏ phiếu" của chúng, thuật toán này có khả năng xử lý tốt các mối quan hệ phi tuyến tính phức tạp trong dữ liệu điện ảnh. Điều này không chỉ tăng độ chính xác mà còn giảm thiểu hiện tượng quá khớp (overfitting).

1. PHẦN 3: PHÂN TÍCH DỮ LIỆU VÀ THIẾT KẾ MÔ HÌNH
   1. Giới thiệu về dữ liệu
      1. Nguồn dữ liệu

Ở đề tài này, bọn em đã đặc biệt tìm kiếm nguồn dữ liệu những bộ phim chiếu rạp tại Việt Nam từ những năm 1990 – 2024. Bộ dữ liệu Movies.csv được đăng tải trên [kaggle.com](https://www.kaggle.com/datasets/trnhhongth/cinema-movies-in-vietnam-1990-2024) thu thập từ nhiều nguồn khác nhau bao gồm Movieek, IMDb và thông qua API TMDb. Nó cung cấp đầy đủ dữ liệu để tiến hành phân tích và áp dụng vào dự án.

* + 1. Các đặc trưng chính

Bộ dự liệu này được thu thập đầy đủ các thông tin cơ bản về một bộ phim đã được phát hành và thống kê lại. Các đặc trưng quan trọng trong tập dữ liệu bao gồm:

* **ID:** Mã định danh duy nhất của bộ phim
* **Title:** Tên phim
* **Overview:** Tóm tắt nội dung phim
* **Revenue:** Doanh thu phòng vé
* **Budget:** Ngân sách sản xuất
* **Runtime:** Thời lượng phim
* **Release Date:** Ngày phát hành
* **Vote Average:** Điểm đánh giá trung bình
* **Vote Count:** Số lượng người đánh giá
* **Genres:** Thể loại phim
* **Production Countries:** Quốc gia sản xuất phim
* **Spoken Languages:** Ngôn ngữ sử dụng trong phim
* **Stars:** Tên các diễn viên chính
  1. Tiền xử lý dữ liệu

Dữ liệu điện ảnh từ nhiều nguồn thường có chất lượng không đồng đều, chứa nhiều giá trị thiếu, trùng lặp hoặc không hợp lệ. Vì vậy, nhóm chúng em đã tiến hành các bước làm sạch dữ liệu để đảm bảo dữ liệu có thể sẵn sàng phục vụ cho phân tích và xây dựng mô hình.

* + 1. Làm sạch dữ liệu và xử lý giá trị thiếu (NaN)

Sau khi xem xét qua bộ dữ liệu, nhận thấy có nhiều giá trị chưa được đồng bộ và đầy đủ. Chúng em đã tiến hành quá trình làm sạch dữ liệu được thực hiện để loại bỏ các giá trị không hợp lệ và xử lý dữ liệu bị thiếu. Cụ thể, các bước được tiến hành:

* Xác định và loại bỏ các hàng có giá trị *Budge*t hoặc *Revenue* bằng 0, vì những giá trị này không phản ánh thực tế ngân sách hoặc doanh thu của bộ phim
* Xử lý các giá trị thiếu trong các cột phân loại như *Genres* và *Production Countries* bằng cách thay thế chúng bằng nhãn "Unknown" hoặc loại bỏ các hàng tương ứng nếu số lượng thiếu không đáng kể
* Kiểm tra và xử lý các giá trị NaN khác trong các cột dữ liệu
  + 1. Chuẩn hóa định dạng dữ liệu
* Cột *Release Date* được chuyển đổi sang định dạng datetime để có thể trích xuất các đặc trưng mới như năm, tháng, và ngày trong tuần
* Các cột chứa dữ liệu dạng phức tạp như *Genres*, *Production Countries* được xử lý để trích xuất các thông tin cần thiết và lưu trữ dưới dạng danh sách các chuỗi
  1. Tạo nhãn thành công

Để tiến hành biến dữ liệu thành một bài toán dự đoán sự thành công của phim chúng em sẽ đặt nó thành một bài toán phân loại nhị phân, chúng ta cần tạo một cột mới trong dữ liệu để xác định xem một bộ phim có được coi là "thành công" hay không. Cột này sẽ đóng vai trò là biến mục tiêu trong mô hình phân loại nhị phân.

Nhưng để định nghĩa “sự thành công” của phim cũng có rất nhiều ý kiến khác nhau, phục thuộc vào nhiều yếu tố khác nhau. Thông thường, các nhà làm phim đều chỉ lấy lợi nhuận làm tiêu chí để đánh giá, nhưng chúng em xét thấy tiêu chí đó chỉ phản ánh được khía cạnh thương mại của phim đó. Trên thực tế, có nhiều bộ phim đạt doanh thu rất cao nhờ chiến dịch quảng bá mạnh mẽ hoặc thương hiệu sẵn có, nhưng lại bị khán giả và giới phê bình đánh giá. Khi đó, việc gán nhãn “thành công” chỉ dựa trên ROI (Return on Investment) sẽ không phản ánh đầy đủ chất lượng của phim.

Ngược lại, cũng có những bộ phim được đánh giá rất cao về nghệ thuật, kịch bản hoặc diễn xuất, nhưng doanh thu phòng vé lại không tương xứng do hạn chế trong phát hành, quảng bá. Nếu chỉ dựa trên rating, ta lại bỏ qua khía cạnh tài chính – vốn là yếu tố then chốt để các nhà sản xuất quyết định đầu tư.

Vì vậy, chúng em đã quyết định lựa chọn tiêu chí kết hợp giữa ROI và Vote Average nhằm đảm bảo tính cân bằng và toàn diện:

* ROI ≥ 1 → đảm bảo phim có thành công tài chính (thu hồi vốn và sinh lợi nhuận
* Vote Average ≥ 6.5 → đảm bảo phim có chất lượng nội dung được khán giả/giới phê bình đánh giá tích cực. Ngưỡng này được chọn vì đây là mức điểm trung bình tốt trên thang đánh giá điểm 10

Trong đó, Vote Average là tiêu chí có sẵn đã được thu thập trong bộ dữ liệu và ROI được tính bằng công thức .

* 1. Phân tích dữ liệu khám phá (EDA)
     1. Thống kê mô tả các đặc trưng số

Thực hiện thống kê mô tả cho các đặc trưng số như Budget, Revenue, và Runtime để nắm bắt các chỉ số trung tâm (trung bình, trung vị) và mức độ phân tán (độ lệch chuẩn). Phân tích này cũng giúp phát hiện các giá trị ngoại lai có thể ảnh hưởng đến mô hình.

* + 1. Trực quan hóa dữ liệu
* Biểu đồ phân phối: Sử dụng biểu đồ histogram để trực quan hóa sự phân bố của Budget và Revenue
* Biểu đồ mối quan hệ: Xây dựng biểu đồ phân tán (scatter plot) giữa Budget và Revenue để phân tích mối tương quan giữa ngân sách và doanh thu
* Biểu đồ tỷ lệ: Biểu đồ cột hoặc biểu đồ tròn được sử dụng để hiển thị tỷ lệ thành công của phim theo các đặc trưng phân loại như Genres và Production Countries
  1. Chuẩn bị dữ liệu cho mô hình
     1. Mã hóa dữ liệu phân loại (One-hot encoding)

Các đặc trưng phân loại như Genres và Production Countries được chuyển đổi sang định dạng số bằng kỹ thuật One-hot encoding. Kỹ thuật này tạo ra các cột nhị phân mới, giúp các thuật toán học máy có thể xử lý được.

* + 1. Chuẩn hóa dữ liệu số

Các đặc trưng số như Budget, Revenue, Runtime, Vote Average và Vote Count được chuẩn hóa bằng phương pháp Min-Max Scaler để đưa các giá trị về cùng một khoảng [0, 1]. Điều này giúp các mô hình học máy hoạt động hiệu quả hơn và tránh hiện tượng các đặc trưng có giá trị lớn hơn lấn át các đặc trưng khác.

* + 1. Phân chia tập dữ liệu

Tập dữ liệu đã được tiền xử lý được chia thành hai phần: tập huấn luyện (training set) chiếm 80% và tập kiểm tra (testing set) chiếm 20%. Tập huấn luyện được sử dụng để xây dựng mô hình, trong khi tập kiểm tra được giữ lại để đánh giá hiệu suất cuối cùng của mô hình một cách khách quan.

* 1. Thiết kế mô hình

Nghiên cứu này lựa chọn và xây dựng ba mô hình phân loại để dự đoán khả năng thành công của phim:

* Logistic Regression: Đây là mô hình phân loại tuyến tính được sử dụng làm mô hình cơ sở (baseline) để so sánh hiệu suất
* Random Forest: Một mô hình học máy ensemble dựa trên việc kết hợp nhiều cây quyết định, giúp giảm thiểu hiện tượng quá khớp và thường cho kết quả chính xác cao

1. PHẦN 4: PHƯƠNG PHÁP THỰC HIỆN VÀ KẾ HOẠCH ĐÁNH GIÁ
   1. Kế hoạch triển khai mô hình

Sau khi hoàn tất quá trình chuẩn bị dữ liệu, chúng em sẽ tiến hành xây dựng và so sánh hai mô hình machine learning: Logistic Regression làm mô hình cơ sở (baseline) và Random Forest làm mô hình chính. Việc lựa chọn hai thuật toán này dựa trên tính chất bổ trợ và phù hợp với trình độ hiện tại của nhóm.

Chúng em sẽ trích một phần bộ dữ liệu để tiến hành huấn luyện (khoảng 240 mẫu) và tập kiểm tra (khoảng 20 - 50 mẫu) để đảm bảo có thể đánh giá khách quan về mô hình triển khai. Việc phân chia sẽ sử dụng phương pháp stratified sampling để duy trì tỷ lệ Success/Failure đồng đều giữa hai tập.

* 1. Phương pháp đánh giá
     1. Các chỉ số đánh giá dự kiến sử dụng

Để đánh giá hiệu suất mô hình một cách toàn diện, chúng em sẽ sử dụng bộ chỉ số sau:

* Confusion Matrix: Cung cấp cái nhìn chi tiết về khả năng phân loại của mô hình, đặc biệt quan trọng trong việc hiểu rõ các loại lỗi. Trong bối cảnh dự đoán thành công phim, việc phân biệt rõ ràng giữa False Positive (đầu tư nhầm) và False Negative (bỏ lỡ cơ hội) có ý nghĩa kinh tế quan trọng
* Accuracy: Đo lường tỷ lệ dự đoán chính xác tổng thể. Mặc dù là chỉ số cơ bản, Accuracy cung cấp cái nhìn tổng quan về hiệu suất mô hình. Chúng em kỳ vọng đạt được Accuracy trên 75% cho cả hai mô hình
* Precision và Recall: Precision đặc biệt quan trọng từ góc độ nhà đầu tư (tránh đầu tư nhầm), trong khi Recall quan trọng từ góc độ nhà sản xuất (không bỏ lỡ cơ hội). Sự cân bằng giữa hai chỉ số này sẽ phản ánh tính thực tế của mô hình
* F1-Score: Là trung bình điều hòa của Precision và Recall, cung cấp đánh giá cân bằng về hiệu suất. Đây sẽ là chỉ số chính để so sánh hai mô hình
  + 1. Kỳ vọng về hiệu suất

Dựa trên báo cáo của những bài nghiên cứu và mô hình tương tự mà chúng em tham khảo được. Kết hợp với sự đặc thù của dữ liệu, chúng em kỳ vọng mô hình của chúng em sẽ có thể đạt được:

* Logistic Regression có thể đạt Accuracy khoảng 70-80% và F1-Score khoảng 65-75%. Mô hình này sẽ hoạt động tốt với các mối quan hệ tuyến tính rõ ràng như tương quan giữa Budget và Success
* Random Forest được kỳ vọng sẽ có hiệu suất cao hơn, có thể đạt Accuracy 80-90% và F1-Score 75-85%. Khả năng xử lý mối quan hệ phi tuyến và tương tác giữa các features sẽ là lợi thế chính
  1. Phương pháp kiểm chứng mô hình dự kiến
     1. Cross-Validation Strategy

Do kích thước dataset tương đối nhỏ (khoảng 300 mẫu), nên chúng em sẽ dùng phương pháp Cross-Validation Strategy (chiến lược kiểm chứng chéo) để đánh giá mô hình bằng cách chia dữ liệu thành nhiều phần (folds), rồi huấn luyện và kiểm tra nhiều lần để có kết quả ổn định và khách quan hơn so với chỉ chia train/test một lần. Chúng em sẽ áp dụng 5-Fold Cross-Validation để tối đa hóa việc sử dụng dữ liệu và đảm bảo tính tin cậy của kết quả. Phương pháp này sẽ giúp:

* Đánh giá độ ổn định của mô hình trên các subset khác nhau
* Phát hiện hiện tượng overfitting nếu có
* Cung cấp confidence interval cho các chỉ số đánh giá
  1. Feature Importance Analysis

Đặc biệt với Random Forest, chúng em sẽ phân tích Feature Importance (tầm quan trọng của đặc trưng) để hiểu rõ yếu tố nào ảnh hưởng quan trọng nhất đến thành công của phim. Kết quả này không chỉ có giá trị khoa học mà còn có ý nghĩa thực tiễn cao cho ngành điện ảnh. Chúng em dự đoán các yếu tố như Budget, Vote\_Count, Runtime, và một số thể loại phim sẽ nằm trong top quan trọng nhất để rút ra xem yếu tố nào là quan trọng đối với khả năng thành công của một bộ phim.

* 1. Kế hoạch phân tích lỗi
     1. Error Analysis Framework

Không có mô hình nào đạt được đến mức hoàn hảo, đặc biệt khi được xây dựng bởi những người mới tiếp cận thực tế như chúng em. Do đó, sau khi thu được kết quả, nhóm sẽ tiến hành phân tích chi tiết các trường hợp mô hình dự đoán sai. Việc này giúp nhận diện những hạn chế, điểm yếu của mô hình:

* False Positives Analysis: Phân tích những bộ phim được dự đoán thành công nhưng thực tế thất bại. Chúng em dự đoán những trường hợp này có thể có ngân sách cao và thể loại hấp dẫn nhưng thất bại do yếu tố chất lượng nội dung hoặc cạnh tranh thị trường
* False Negatives Analysis: Nghiên cứu những bộ phim bị đánh giá thấp nhưng thực sự thành công. Đây có thể là những bộ phim có ngân sách thấp nhưng viral nhờ content độc đáo hoặc marketing sáng tạo
  + 1. Business Impact Assessment

Nhóm sẽ xem xét tác động kinh tế của từng loại lỗi dự đoán. Cụ thể, các trường hợp False Positive có thể dẫn đến tổn thất trực tiếp do ra quyết định đầu tư sai lầm vào dự án không thành công. Ngược lại, False Negative gây ra tổn thất cơ hội (opportunity cost) khi bỏ lỡ những dự án tiềm năng đáng lẽ có thể mang lại giá trị kinh tế. Việc phân tích này giúp làm rõ rủi ro gắn liền với từng loại sai sót, từ đó định hướng lựa chọn mô hình phù hợp với mục tiêu thực tiễn.

* 1. Thách thức dự kiến và giải pháp
     1. Thách thức về dữ liệu

Về dữ liệu, tình trạng mất cân bằng giữa hai lớp Success và Failure có thể khiến mô hình dự đoán thiên lệch, vì vậy chúng em sẽ áp dụng các kỹ thuật như SMOTE hoặc class weighting để cân bằng. Quy mô dữ liệu nhỏ (khoảng 300 mẫu) cũng làm gia tăng nguy cơ overfitting; do đó, nhóm sẽ kết hợp regularization với cross-validation nhằm nâng cao khả năng tổng quát hóa. Ngoài ra, việc tồn tại các giá trị khuyết trong một số đặc trưng là điều khó tránh khỏi, và chúng em dự kiến sẽ xử lý bằng các phương pháp imputation hoặc bổ sung biến chỉ báo “is\_missing” để giảm thiểu mất mát thông tin.

* + 1. Thách thức về mô hình

Về mô hình, một rủi ro khác là cả Logistic Regression và Random Forest có thể không đạt kết quả như kỳ vọng; khi đó, nhóm sẽ triển khai hyperparameter tuning và đẩy mạnh feature engineering để cải thiện hiệu suất. Cuối cùng, bài toán cân bằng giữa hiệu năng và khả năng giải thích cũng là một thách thức, bởi Random Forest thường cho kết quả tốt hơn nhưng khó minh bạch hơn Logistic Regression. Nhóm sẽ cân nhắc lựa chọn sao cho vừa đảm bảo tính chính xác, vừa giữ được mức độ dễ hiểu cần thiết cho ứng dụng thực tiễn.

1. PHẦN 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN
   1. Tổng kết

Đề tài “Dự đoán độ thành công của phim” bước đầu đã chứng minh được tính khả thi. Nhóm chúng em đã xây dựng được một phương pháp nghiên cứu rõ ràng, từ việc định nghĩa tiêu chí “thành công”, tiền xử lý dữ liệu, đến lựa chọn mô hình và cách đánh giá. Quan trọng hơn, nhóm đã chuyển đổi một vấn đề mang tính kinh doanh trừu tượng thành một bài toán kỹ thuật có thể giải quyết bằng học máy.

* 1. Ý nghĩa và tác động thực tiễn

Kết quả nghiên cứu dự kiến mang lại ý nghĩa cho nhiều đối tượng:

* Nhà sản xuất và đầu tư: có thêm công cụ hỗ trợ ra quyết định dựa trên dữ liệu thay vì chỉ dựa vào kinh nghiệm
* Nhà phát hành và marketing: có cơ sở để lựa chọn thời điểm phát hành và phân bổ ngân sách hiệu quả hơn
* Giới nghiên cứu: cung cấp một khung nghiên cứu mẫu để áp dụng Machine Learning trong các ngành sáng tạo khác như âm nhạc hay game
  1. Nhận thức về hạn chế

Nhóm nhận thức rằng nghiên cứu còn nhiều giới hạn:

* Thiếu các yếu tố định tính quan trọng như chất lượng kịch bản hay cạnh tranh thị trường
* Định nghĩa “thành công” vẫn còn giới hạn, chủ yếu dựa vào ROI và đánh giá người xem, chưa phản ánh hết giá trị dài hạn
* Kích thước dữ liệu nhỏ, nguy cơ overfitting cao
* Thị trường điện ảnh thay đổi nhanh, mô hình có thể mất tính chính xác trong tương lai
  1. Lời kết

Qua quá trình nghiên cứu, nhóm nhận thấy rằng việc ứng dụng Machine Learning trong dự đoán mức độ thành công của phim không chỉ mang ý nghĩa về mặt học thuật, mà còn có tiềm năng đóng góp thực tiễn cho ngành công nghiệp điện ảnh. Trong bối cảnh thị trường điện ảnh luôn tiềm ẩn rủi ro và đòi hỏi các quyết định đầu tư lớn, việc khai thác dữ liệu và mô hình dự đoán có thể trở thành một công cụ hỗ trợ ra quyết định hiệu quả hơn so với việc chỉ dựa vào kinh nghiệm hoặc trực giác.

Nghiên cứu bước đầu đã cho thấy sự kết hợp giữa khoa học dữ liệu và lĩnh vực điện ảnh có thể mang lại giá trị thiết thực. Thay vì chỉ tập trung vào những hãng phim lớn có khả năng tiếp cận nghiên cứu thị trường chuyên sâu, mô hình này có thể được mở rộng để hỗ trợ các nhà làm phim độc lập hoặc các nhà đầu tư nhỏ, giúp họ có thêm cơ sở dữ liệu đáng tin cậy trong quá trình ra quyết định.

Nhìn chung, nghiên cứu góp phần đặt nền móng cho việc ứng dụng các phương pháp phân tích dữ liệu hiện đại vào ngành giải trí, đồng thời hướng tới mục tiêu xây dựng một hệ thống hỗ trợ ra quyết định có cơ sở khoa học, minh bạch và bền vững hơn cho lĩnh vực điện ảnh.

TÀI LIỆU THAM KHẢO

1. Tham khảo paper và sách nền tảng

* Bishop, C. M. Pattern Recognition and Machine Learning. Springer, 2006
* Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O’Reilly, 2019
* Github: “BigData\_Du-Doan-Xu-Huong-Phim”

1. Tham khảo từ dữ liệu:

* Kaggle dataset: “Cinema Movies in VietNam 1990 - 2024”