Báo Cáo Tóm Tắt Tuần 1-4: Dự Án Dự Đoán Thành Công Phim

Nhóm 04 - Khoa học Dữ liệu HUMG

Giới Thiệu Tổng Quan

Dự án "Dự Đoán Độ Thành Công Phim Chiếu Rạp Tại Việt Nam" nhằm xây dựng một công cụ thông minh để dự đoán thành công của phim dựa trên các yếu tố như ngân sách, thể loại, diễn viên và quốc gia sản xuất. Báo cáo này tóm tắt những gì nhóm đã thực hiện trong 4 tuần đầu, từ thu thập thông tin gốc đến hoàn thành việc tạo ra các đặc điểm phân tích, cùng với các kết quả, bất cập và hướng giải quyết.

Tuần 1: Khởi Động & Thu Thập Dữ Liệu

Những gì đã làm

- **Thu thập bộ thông tin:** Nhận file `Movies.csv` chứa 2194 bộ phim với 17 thông tin cơ bản (tên phim, ngân sách, doanh thu, thể loại, diễn viên, đạo diễn, ngày phát hành, điểm đánh giá)
- **Phân tích thông tin ban đầu:** Sử dụng công cụ để kiểm tra cấu trúc, kiểu thông tin, và phát hiện các vấn đề tiềm ẩn
- **Thống nhất tiêu chí thành công:** Họp nhóm và giảng viên để định nghĩa phim "thành công" = (ROI ≥ 1.0) VÀ (Vote Average ≥ 6.5)
- **Thiết lập môi trường:** Tạo cấu trúc thư mục dự án, cài đặt các công cụ và thư viện cần thiết

Kết quả đạt được

- **Bộ thông tin sẵn sàng:** File `Movies.csv` với 2194 hàng × 17 cột, không có lỗi đọc file
- **Hiểu rõ thông tin:** Phát hiện 1173 phim có ngân sách/doanh thu = 0, định dạng thể loại phức tạp, 953 phim thiếu thông tin diễn viên/đạo diễn

- **Tiêu chí rõ ràng:** Toàn nhóm thống nhất định nghĩa thành công dựa trên cả yếu tố tài chính (ROI) và chất lượng (điểm đánh giá)
- **Môi trường ổn định:** Cấu trúc thư mục có tổ chức, danh sách thư viện sẵn sàng

Bất cập gặp phải

- **Thiếu kinh nghiệm quản lý nhóm:** Cuộc họp đầu tiên kéo dài, khó đạt được sự thống nhất
- **Thông tin phức tạp:** Thể loại ở dạng danh sách văn bản "['Action', 'Drama']" khó xử lý
- **Chưa rõ hướng xử lý:** Chưa biết cách xử lý 1173 hàng có giá trị 0

Hướng giải quyết

- **Áp dụng phương pháp linh hoạt:** Thiết lập chương trình họp rõ ràng cho các cuộc họp tiếp theo
- **Ghi chú chi tiết:** Lưu trữ tất cả quyết định để tham khảo sau
- **Lên kế hoạch tiền xử lý:** Chuẩn bị chiến lược cho tuần 2

Tuần 2: Làm Sạch Thông Tin Cơ Bản

Những gì đã làm

- **Xử lý giá trị 0:** Loại bỏ 1173 hàng có ngân sách hoặc doanh thu = 0 vì không thể tính ROI chính xác
- **Xử lý thông tin thiếu:** Điền 'Unknown' cho cột văn bản (đạo diễn, diễn viên), điền giá trị trung bình cho cột số (thời lượng phim)
- **Chuẩn hóa ngày tháng:** Chuyển ngày phát hành sang định dạng ngày tháng chuẩn
- **Tạo công cụ tự động:** Viết `cleandata.py` để có thể lặp lại quá trình làm sạch

Kết quả đạt được

- **Bộ thông tin sạch:** File `clean_movies.csv` với 1020 hàng (giảm 53% nhưng chất lượng cao)
- **Không còn thông tin thiếu:** 100% thông tin đầy đủ trong các cột quan trọng

- **Định dạng thống nhất:** Ngày phát hành ở dạng ngày tháng chuẩn, có thể trích xuất các đặc điểm thời gian
- **Quy trình có thể lặp lại:** Công cụ `cleandata.py` có thể chạy lại với bộ thông tin khác

Lưu ý quan trọng

- **Chất lượng > Số lượng:** Quyết định xóa 53% thông tin để đảm bảo tính chính xác của ROI
- **Chiến lược điền thiếu:** 'Unknown' cho văn bản giúp công cụ biết đây là thông tin thiếu
- **Xử lý lỗi nhẹ nhàng:** Sử dụng cách xử lý lỗi để xử lý ngày tháng bị lỗi

Bất cập gặp phải

- **Tranh cãi trong nhóm:** Một số thành viên lo ngại mất quá nhiều thông tin
- **Hiệu suất chậm:** Công cụ ban đầu chạy >30 giây do vòng lặp không tối ưu
- **Lỗi phân tích ngày tháng:** 1 giá trị không phân tích được

Hướng giải quyết

- **Thử nghiệm thực tế:** So sánh kết quả khi điền vs xóa thông tin, chứng minh xóa là đúng
- **Tối ưu mã nguồn:** Thay vòng lặp bằng các phép toán vector, giảm thời gian xuống <5 giây
- **Chấp nhận mất mát tối thiểu:** 1 giá trị lỗi (<0.1%) không đáng kể, giữ nguyên

---

Tuần 3: Tạo Nhãn Thành Công & Khám Phá Thông Tin Cơ Bản

Những gì đã làm

- **Tính ROI:** ROI = Doanh thu / Ngân sách cho 1020 phim còn lại
- **Tạo nhãn thành công:** Áp dụng công thức (ROI ≥ 1.0) VÀ (điểm đánh giá ≥ 6.5)
- **Trích xuất đặc điểm thời gian:** Tạo năm phát hành, tháng phát hành, ngày trong tuần từ ngày phát hành
- **Vẽ biểu đồ khám phá:** Vẽ 3 biểu đồ chính - phân bố ROI, ngân sách theo thành công, tỷ

lệ thành công theo thể loại

Kết quả đạt được

- **Biến mục tiêu cân bằng:** 514 phim thành công vs 506 phim thất bại (tỷ lệ 1.01:1)
- **Bộ thông tin với nhãn:** File `clean_movies_with_labels.csv` với 1020 hàng × 25 cột
- **Hiểu biết quan trọng:** Hành động/Phiêu lưu có tỷ lệ thành công cao nhất, phim gần đây thành công hơn
- **Đặc điểm thời gian:** 3 cột thời gian mới để phân tích xu hướng theo mùa

Lưu ý quan trọng

- **Cân bằng lớp tuyệt vời:** Không cần các kỹ thuật cân bằng lại thông tin đặc biệt
- **Phân bố ROI:** Đa số phim có ROI 0.2-0.8, chỉ 40% phim thực sự lãi
- **Hiểu biết theo mùa:** Mùa hè và lễ hội có xu hướng thành công cao hơn

Bất cập gặp phải

- **Ngưỡng điểm đánh giá:** Tranh luận về 6.5 có quá cao so với phim Việt Nam
- **Giá trị ngoại lệ trong ROI:** Một số phim có ROI cực cao làm lệch phân bố
- **Phân tích thể loại phức tạp:** Khó so sánh vì mỗi phim có nhiều thể loại

Hướng giải quyết

- **Thử nhiều ngưỡng:** Thử 6.0, 6.5, 7.0 và chọn 6.5 vì cân bằng nhất
- **Trực quan hóa theo thang logarit:** Sử dụng thang log cho ngân sách để thấy mẫu rõ hơn
- **Tập trung thể loại hàng đầu:** Chỉ phân tích top 10 thể loại phổ biến nhất

---

Tuần 4: Tạo Đặc Điểm Phân Tích

Những gì đã làm

- **Mở rộng đặc điểm thời gian:** Thêm quý phát hành, mùa lễ hội (5 đặc điểm từ ngày phát

hành)
- **Xử lý thời lượng phim:** Nhóm thành 5 loại, tạo thời lượng tính bằng phút và giờ
- **Đếm diễn viên:** Phân tích cột diễn viên để đếm số diễn viên chính
- **Mã hóa phân loại:** Mã hóa nhị phân cho top 15 thể loại và top 10 quốc gia
- **Biến đổi số:** Biến đổi logarit cho ngân sách/doanh thu, cắt bớt các giá trị ngoại lệ ROI
- **Đặc điểm tương tác:** Tạo ngân sách theo năm, ROI kết hợp điểm, tương tác diễn viên-thể loại

Kết quả đạt được

- **Bộ thông tin sẵn sàng:** File `clean_movies_features.csv` với 1020 hàng × 65 cột
- **Đa dạng đặc điểm:** Hơn 40 đặc điểm số bao gồm thời gian, phân loại, tương tác
- **Phân tích mạnh mẽ:** Xử lý thành công 100% thông tin thể loại/quốc gia phức tạp
- **Chất lượng được xác thực:** Không có thông tin thiếu, phân bố hợp lý, sẵn sàng cho thuật toán học máy

Lưu ý quan trọng

- **Kiểm soát chiều dữ liệu:** 65 đặc điểm với 1020 mẫu = tỷ lệ 15:1 (chấp nhận được)
- **Cần chọn lọc đặc điểm:** Sẽ dùng mức độ quan trọng từ rừng ngẫu nhiên ở tuần 5
- **Dựa trên logic kinh doanh:** Mỗi đặc điểm đều có lý do từ kiến thức chuyên môn

Bất cập gặp phải

- **Phân tích chuỗi phức tạp:** Nhiều định dạng cho thể loại/quốc gia gây khó khăn
- **Nguy cơ chiều dữ liệu:** 65 đặc điểm có thể gây quá khớp
- **Đặc điểm tương tác đáng ngờ:** Một số tương tác không có logic kinh doanh rõ ràng
- **Sử dụng bộ nhớ:** Xử lý 65 đặc điểm cần tối ưu bộ nhớ

Hướng giải quyết

- **Phương pháp phân tích lại:** Thử phân tích JSON trước, dự phòng bằng tách biểu thức chính quy
- **Kế hoạch chọn đặc điểm:** Mức độ quan trọng từ rừng ngẫu nhiên cho tuần 5
- **Xác thực tương tác:** Sẽ kiểm tra tương quan với mục tiêu, loại bỏ nếu không có ý nghĩa
- **Tối ưu bộ nhớ:** Sử dụng kiểu dữ liệu phù hợp (int8 cho đặc điểm nhị phân)

Đánh Giá Tiến Độ

- **Hoàn thành:** 100% Tuần 1-4 (4/10 tuần), đúng kế hoạch. Quy trình từ thông tin thô → đặc điểm sạch, không bị trễ.
- **Điểm mạnh:** Chất lượng thông tin cao, trực quan hóa tốt, tài liệu chi tiết.
- **Điểm yếu:** Mất thời gian cho tiền xử lý (40% công sức), cần tối ưu tự động hóa.

Bài Học Rút Ra

- **Tiền xử lý quan trọng:** 80% thành công của học máy nằm ở chất lượng thông tin; xử lý thiếu/ngoại lệ sớm tránh lỗi sau.
- **Kiến thức chuyên môn:** Hiểu về phim (ROI, thể loại) giúp tạo đặc điểm hiệu quả hơn ngẫu nhiên.
- **Hợp tác:** Git/PR giúp tránh xung đột, họp định kỳ tăng hiệu quả.
- **Trực quan hóa:** Biểu đồ không chỉ đẹp, mà phát hiện hiểu biết (vd: tác động thể loại).

Định Hướng Cho Giai Đoạn Tiếp Theo (Tuần 5-8)

- **Tuần 5:** Mô hình cơ bản (hồi quy logistic: dự kiến độ chính xác ~85%, F1 ~85%) và chính (rừng ngẫu nhiên: mục tiêu độ chính xác >95%, F1 >95%), đánh giá với các thước đo chính: độ chính xác, F1-Score, ROC-AUC.
- **Tuần 6:** Mức độ quan trọng đặc điểm từ RF, so sánh các mô hình.
- **Tuần 7:** Xử lý mất cân bằng nếu cần, phân tích lỗi.
- **Tuần 8:** Xác thực chéo, ổn định mô hình.
- **Mục tiêu:** Hoàn thành mô hình cuối với độ chính xác >95%, F1 >95%, ROC-AUC >0.95, vượt cơ bản logistic ~85%.
- **Rủi ro:** Quá khớp với nhiều đặc điểm; giải pháp: CV, điều chuẩn.
- **Kế hoạch:** Tăng tự động hóa (scripts), thử nghiệm trên thông tin chưa thấy, đánh giá hàng tuần.

Quản Lý Nhóm: Phân Công Công Việc Tuần 1-4

Thành Viên	Vai Trò Chính	Công Việc Tuần 1	Công Việc Tuần 2	Công Việc Tuần 3	Công Việc Tuần 4
Khổng Thị Hoà	Trưởng nhóm	Họp nhóm, thống nhất tiêu chí	Thiết lập môi trường, Xử lý thiếu, làm sạch thông tin, Lưu bộ thông tin sạch	Khám phá thông tin	Họp nhóm đánh giá đặc điểm
Phan Văn Huy	Khám phá/ Trực quan	Phân tích thông tin gốc	Điền thiếu, chuyển ngày tháng	Vẽ biểu đồ, thống kê, Tổng hợp khám phá	Phân tích chuỗi, mã hóa
Đinh Ngọc Khuê	Chuẩn bị mô hình/ Tài liệu	Viết báo cáo tuần 1, Ghi chú quyết định		Tạo nhãn, Tính ROI, thêm đặc điểm	Tạo đặc điểm, Tạo đặc điểm kết hợp, Kiểm tra khả năng lặp lại

Người đảm nhận chính từng tuần: Tuần 2 - Khổng Thị Hoà, Tuần 3 - Phan Văn Huy, Tuần 4 - Đinh Ngọc Khuê.

Kế Hoạch Tương Lai (Tuần 5-10)

Tuần 5: Chia Thông Tin & Xác Thực Chéo

Mục tiêu: Chia bộ thông tin huấn luyện/kiểm tra và thiết lập xác thực chéo K-fold

- **Kết quả:** Chia huấn luyện/kiểm tra 80/20, thiết lập xác thực chéo 5-fold, thước đo cơ bản
- **Thời gian:** 7 ngày, đã có nền tảng từ tuần 1-4

Tuần 6-8: Huấn Luyện & Đánh Giá Mô Hình

Phương pháp: So sánh 3 thuật toán chính

- **Hồi quy Logistic:** Mô hình cơ bản có thể giải thích
- **Rừng Ngẫu Nhiên:** Phương pháp tập hợp dựa trên cây
- **Máy Vector Hỗ Trợ:** Phân loại phi tuyến
- **Đánh giá:** Độ chính xác, Precision, Recall, F1-score, ROC-AUC

Tuần 9: Lựa Chọn & Tối Ưu Mô Hình

Tập trung: Điều chỉnh siêu tham số và lựa chọn đặc điểm

- **Tìm kiếm lưới:** Tối ưu siêu tham số cho mô hình tốt nhất
- **Mức độ quan trọng đặc điểm:** Chọn đặc điểm quan trọng nhất
- **Mô hình cuối:** Bộ phân loại sẵn sàng sản xuất

Tuần 10: Triển Khai & Tài Liệu

Kết quả: Dự án hoàn chỉnh với báo cáo cuối

- **Triển khai mô hình:** Ứng dụng Streamlit cho demo
- **Báo cáo cuối:** Tài liệu toàn diện
- **Thuyết trình:** Thuyết trình nhóm 15 phút

Tổng Kết 4 Tuần Đầu

Thành Tựu Chính

Bộ thông tin chất lượng cao: 1020 hàng × 65 đặc điểm, sẵn sàng cho học máy
Phối hợp nhóm: Phân công rõ ràng, không trễ hạn
Nền tảng kỹ thuật: Mã nguồn sạch, có thể lặp lại, tài liệu tốt

Hiểu biết lĩnh vực: Hiểu rõ logic kinh doanh và hiểu biết ngành công nghiệp phim

Bài Học Rút Ra

- **Chất lượng > Số lượng:** Xóa 53% thông tin nhưng đảm bảo chính xác của ROI
- **Giao tiếp nhóm:** Gặp gỡ hàng ngày ngắn hiệu quả hơn cuộc họp hàng tuần dài
- **Kiến thức lĩnh vực quan trọng:** Hiểu các yếu tố thành công phim giúp tạo đặc điểm tốt hơn
- **Tài liệu quan trọng:** Ghi chép chi tiết giúp truy vết quyết định

Mức Độ Tin Tưởng cho Tuần 5-10

Tin tưởng cao (90%): Quy trình thông tin vững chắc, nhóm đã đồng bộ tốt
Tin tưởng trung bình (70%): Hiệu suất mô hình - bộ thông tin nhỏ có thể hạn chế độ chính xác
Kế hoạch giảm thiểu: Sẽ tập trung vào tạo thêm đặc điểm và phương pháp tập hợp nếu mô hình đơn không đủ tốt

---

Phụ Lục: Thông Số Kỹ Thuật

Tóm Tắt Quy Trình Thông Tin

```
Movies.csv (2194×17)
→ cleandata.py → clean_movies.csv (1020×17)
→ Khám phá + gắn nhãn → clean_movies_with_labels.csv (1020×25)
→ feature_engineering.ipynb → clean_movies_features.csv (1020×65)
```

Các Đặc Điểm Chính Được Tạo (tổng 65)

- **Đặc điểm thời gian (5):** năm, tháng, ngày trong tuần, quý, mùa lễ hội
- **Đặc điểm thời lượng (3):** loại, phút, giờ
- **Đặc điểm diễn viên (1):** số diễn viên chính
- **Mã hóa thể loại (15):** Cột nhị phân cho thể loại hàng đầu
- **Mã hóa quốc gia (10):** Cột nhị phân cho quốc gia hàng đầu

- **Đặc điểm tài chính (4):** budget_log, revenue_log, roi, roi_clipped
- **Đặc điểm tương tác (3):** budget_per_year, roi_vs_vote, cast_genre_interaction
- **Đặc điểm gốc (24):** Giữ lại từ bộ thông tin sạch

Xác Thực Tiêu Chí Thành Công

- **ROI ≥ 1.0:** Phim sinh lời (góc độ kinh doanh)
- **Vote Average ≥ 6.5:** Chất lượng trên trung bình (góc độ khán giả)
- **Kết quả:** 514 thành công vs 506 thất bại (lớp cân bằng)

Kết Luận: Tuần 1-4 đã xây dựng nền tảng vững chắc, với bộ thông tin sạch và đặc điểm phong phú. Nhóm sẵn sàng cho mô hình, kỳ vọng mô hình tốt cho dự đoán phim. Nếu cần chỉnh sửa, liên hệ nhóm.