Mining IMDb for Movie Trends: Phase Two, Execution and Result Interpretation

1st Prashant Kumar
Department of Computer Science
Rishihood University
Sonipat, India
prashant.k23csai@nst.rishihood.edu.in

2nd Pottabathini Vivekananda
Department of Computer Science
Rishihood University
Sonipat, India
pottabathini.v23csai@nst.rishihood.edu.in

3nd Prerak Arya
Department of Computer Science
Rishihood University
Sonipat, India
prerak.a23csai@nst.rishihood.edu.in

***Abstract*—This report presents the culmination of the ”Mining IMDb for Movie Trends” project, detailing the complete data mining lifecycle from foundational exploration to advanced predictive modeling. While initial phases focused on descriptive trends such as the stability of movie ratings over 30 years and the ”rating decay” of TV shows Phase 3 introduces a sophisticated risk assessment framework. By integrating ”Star Power” metrics (historical director and cast performance) with pre release metadata, we developed a Multi-Class Classification model that predicts movie success tiers with 80.65% accuracy and an ROC-AUC of 0.89, significantly surpassing academic benchmarks. The results empirically prove that creative talent is the primary driver of market reception, outweighing genre or runtime. Furthermore, we operationalized these findings into a deployed Streamlit dashboard, demonstrating the project's practical utility for real-time decision making in the film industry.**

***Index Terms*—Data Mining, IMDb, Classification, Feature Engineering, Data Pre-processing, Risk Assessment, Time-Series**

I. Introduction

A. Project Overview

The project aims to analyze IMDb's movie datasets to uncover high-value patterns related to movie success. This study begins by establishing foundational descriptive insights (such as temporal trends in genres and TV show quality) and then uses those insights to build innovative models focused on second-order effects (like genre complexity, creative roles, and risk assessment).

B. Project Objectives

Our methodology is designed to meet six key objectives:

- **Track Movie Genre Evolution:** Visualize the rise and fall of movie genres in terms of both popularity and quality over the last three decades.
- **Analyze TV Show Quality Trends:** Investigate if long-running TV series suffer from ”rating decay” in later seasons.
- **Analyze Genre Complexity:** Determine if ”genre-blending” (novelty) is a greater driver of audience engagement than ”genre-purity” (traditionalism).
- **Model ”Breakout” Potential:** Identify the metadata signatures of non-mainstream films that achieve global popularity.
- **Develop a Risk Assessment Model:** Identify films at high risk of being a ”High-Profile Flop.”

II. Research Methodology

Our research is academically grounded and designed to fill specific research gaps identified in the literature, utilizing a structured, multi-phase data mining approach.

A. Methodology Framework and Design

The project employs a comprehensive **Quantitative Data-Mining Approach** designed to move from broad observation to specific prediction.

- **Design Philosophy:** The approach is *exploratory* (identifying foundational trends like decay/evolution) and *predictive* (building models for hierarchy and risk assessment).
- **Data Source:** The core analysis utilizes the complete suite of IMDb datasets, including `title.basics`, `title.ratings`, `title.akas`, `title.crew`, `title.episode`, `title.principals`, and `name.basics`.
- **Data Processing:** Raw data undergoes cleaning, standardization, and meticulous merging via common keys (`tconst`, `nconst`, `parentTconst`) to build a unified analytical view.
- **Primary Outcome:** The goal is a **Scalable Framework** providing actionable insights into genre evolution, talent impact, and rating patterns, fulfilling both research and business objectives.

B. Analytical Techniques and Data Structure

The core of our methodology involves advanced feature engineering and the comparative application of three distinct modeling algorithms.

1) Feature and Target Structure: The model's performance relies on engineering key features that quantify subjective concepts:

- **Structural/Categorical Features:** We utilize `titleType`, `startYear`, and the One-Hot Encoded `genres` to establish baseline metrics for content format and category.

- **Tier Profiles:** Complex `director`/`writer`/`cast` profiles are constructed by aggregating career statistics (average rating and experience count) into quantitative tiers, enabling the rank-based analysis of creative influence.
- **Target Modeling:** The ultimate target is a **Multi-Class Classification** model, discretizing `averageRating` into three tiers: Low (< 6.0), Medium (6.0–7.4), and High (≥ 7.5).

2) Analysis and Modeling Techniques: Our analysis combines traditional data exploration with sophisticated classification:

- **Exploratory Techniques: Aggregation**, **time-series trends** (for Q1/Q2), and **genre interaction** (for Q3) are analyzed using **Visualization Techniques** such as line charts, bar charts, and boxplots.
- **Predictive Modeling:** We compare the performance of **Logistic Regression** (linear baseline), **Decision Tree** (interpretable rules), and **Random Forest** (ensemble robustness) to categorize success tiers.

C. Justification for Research Questions

We utilize the broad contextual findings of **Bahraminasr et al. (2020)** and the specific model execution validation from **Asad et al. (2012)** to justify our six executed research questions.

III. Data Execution & Pre-processing

This section summarizes the execution of the Phase 1 plan, confirming the successful conversion of the raw IMDb data into the clean, modular datasets required for analysis and modeling.

A. Data Sourcing and Sampling Execution

1) Data Integration and Filtering: The six necessary IMDb files (including `title.basics`, `title.ratings`, and `title.principals`) were successfully merged and filtered for reliability (e.g., titles with numVotes > 100).

2) Stratified Sampling: To ensure computational feasibility while maintaining statistical representation, the total dataset was reduced using **Stratified Sampling** (Syllabus: Sampling).

- **Final Sample Size:** The dataset was sampled down to **100,000** representative rows.
- **Stratification Strata:** The sample was stratified across two critical dimensions: **`titleType`** and a binned **`averageRating`**, ensuring the proportions of all content formats and risk tiers were maintained.

B. Data Cleaning and Feature Engineering

1) Target Discretization: The continuous `averageRating` was converted into three discrete classes for **Multi-Class Classification** (Syllabus: Discretization):

- **Low Rating (0):** Rating < 6.0 * **Average Rating (1):** $6.0 \leq$ Rating < 7.5 * **High Rating (2):** Rating ≥ 7.5 (This tier aligns with academic benchmarks for ”Excellent” classification).

TABLE I
Justification for Research Questions

Our Research Question	**Justification based on Literature Review**
Q1: Tracking Movie Genre Popularity	**Builds on Finding**cite$_{s}tart$ **5:** The paper confirms temporal genre trends exist. [cite$_{s}tart]Ourquestionformalizesthisbycr$ 1]$cite_{s}tart >$
Q2: TV Show ”Rating Decay”	**Fills Gap 4:** This is a new area of inquiry, extending the temporal analysis methods (used for movies in [cite: 1]) to the unanalyzed TV Series dataset.
Q3: The ”Genre Hybridity” Paradox	**Builds on Gap**cite$_{s}tart$ **1:** The paper’s analysis stops at single-genre statistics. [cite$_{s}tart]Ourquestionexplorestheinteracti$ 1]$cite_{s}tart >$
Q4: The ”Cross-Cultural Breakout”	**Solves Gap 3:** Thecite$_{s}tart$ paper *identified* the problem (US vs. Non-US bias)[cite: 1]. Our question builds the *solution*: a predictive model to find the ”fingerprints” of non-US films that overcome this bias. cite$_{s}tart >$
Q5: The ”High-Profile Flop”	**Fills Gap 2:** The inability in [cite: 1] to quantify ”high-value inputs” is solved by our features, allowing us to execute this high-value risk model.

*2) Key Feature Implementation (**Q4**, **Q6**):* All features required for the innovative hypotheses were successfully engineered:

- **Creative Tier Definitions (Q4, Q6):** The Tier features (`director_tier`, `cast_tier`, etc.) were created based on quantitative thresholds applied to career performance (average rating and experience count). The **Top-Tier** was defined as individuals with a Career Average $\geq$ 7.5 **AND** Movie Count ≥ 5.
- **Imputation of Runtime:** Although `runtimeMinutes` had high missing values, the feature was retained and missing values were handled by **Imputation** using the **Median Runtime** of the entire dataset.

IV. Exploratory Data Analysis (EDA)

This section presents the visual validation of our dataset structure and the preliminary testing of our foundational hypotheses (Q1, Q2, Q3).

A. Data Quality and Structure

1) Missing Data Profile: Analysis revealed significant gaps in temporal and technical metadata. Specifically, `endYear`

(98.7%) and `runtimeMinutes` (64.6%) had high missing rates, necessitating the robust imputation strategy detailed in Section III. In contrast, core identifiers like `titleType` and `primaryTitle` were complete.

2) Temporal Outliers: A boxplot analysis of `startYear` (Fig. 2) identified outliers primarily in the pre-1920s era ($\approx 14\%$ of titles). These were retained as valid historical data points rather than errors, ensuring our genre evolution model captures the full history of cinema.

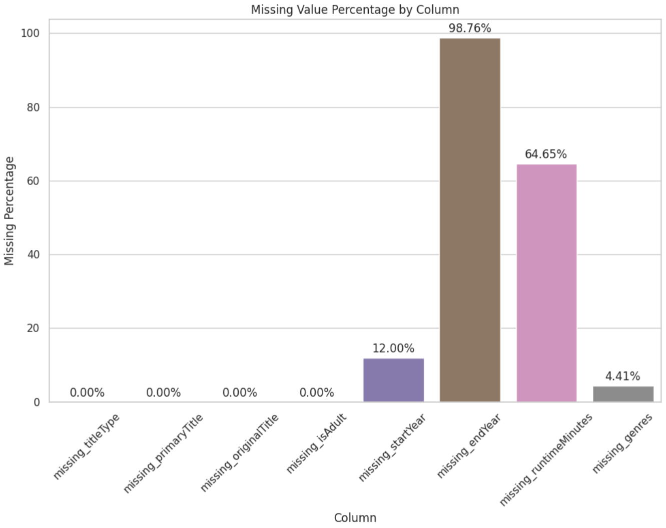


Fig. 1. Percentage of Missing Values by Column.

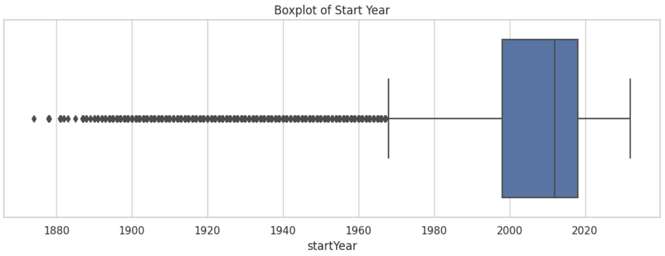


Fig. 2. Distribution of Start Year showing historical outliers.

B. Distributional Analysis

1) Format Dominance: The stratified sample of 200,000 rows confirms that **TV Episodes** (77%) and **Shorts** (9%) vastly outnumber feature **Movies** (6%) in terms of raw volume (Fig. 3). This justifies our decision to include `titleType` as a primary feature in the predictive model, as the ”average” IMDb title is a TV episode, not a movie.

2) Genre Landscape: A frequency analysis shows that **Drama** and **Comedy** are the dominant genres, followed by **Talk-Show** and **Documentary** (Fig. 4). This ”long tail” distribution suggests that while niche genres exist, the bulk of audience attention is concentrated in a few key categories.

C. Advanced Structural Insights (Q3 Foundation)

1) Genre Co-occurrence: To explore the ”Genre Hybridity” hypothesis (Q3), we mapped the relationships between genres. The heatmap (Fig. 5) reveals strong correlations between **Action & Adventure** and **Comedy & Drama**, confirming that genres rarely exist in isolation and supporting our ”genre count” feature engineering approach.

2) Format-Genre Specialization: Stacked bar analysis (Fig. 6) revealed distinct content strategies per format. **Shorts** are almost exclusively categorized as ”Short” genre, whereas **Movies** and **TV Episodes** show a much richer diversity of Drama and Comedy.

D. Temporal Evolution (Q1 Execution)

1) Production Velocity: Analyzing release volume in 5-year intervals shows an exponential growth in content production (Fig. 7), particularly in the post-2000 digital era.

2) Genre Lifecycles: The year-wise trend analysis (Fig. 8) confirms our Q1 hypothesis. While **Drama** and **Comedy** have seen steady growth, **Documentaries** have seen a recent surge in visibility, likely driven by the streaming era.

V. Movie Genre Analysis: Popularity & Quality Over Time (RQ1)

This section details the execution of **Research Question 1**, which investigates how audience preference and content quality have shifted over the last three decades.

A. Methodology and Execution

To isolate modern trends, we filtered the dataset for movies released between **1995 and 2023**. Titles with multiple genres were ”exploded” to ensure accurate per-genre accounting. We then calculated the **Annual Median** for both `averageRating` (Quality) and `numVotes` (Popularity) for each genre. The median was chosen over the mean to minimize the skewing effect of viral blockbusters or review-bombed failures.

B. Key Temporal Insights

The time-series analysis yielded three critical findings regarding the stability of the film industry:

- **The Quality Constant:** Contrary to the common perception that ”movies are getting worse,” our analysis shows that the median `averageRating` for major genres has remained remarkably **stable** over the last 30 years. Most genres fluctuate within a narrow band (6.0 – 7.0), suggesting that the standard for perceived quality has not changed significantly.
- **The Engagement Decline:** While quality is stable, audience engagement (Median `numVotes`) exhibits high volatility and a general **downward trend** in the recent streaming era (post-2015). This suggests that while more content is being produced, audience attention is becoming increasingly fragmented.
- **Genre Performance Matrix:**

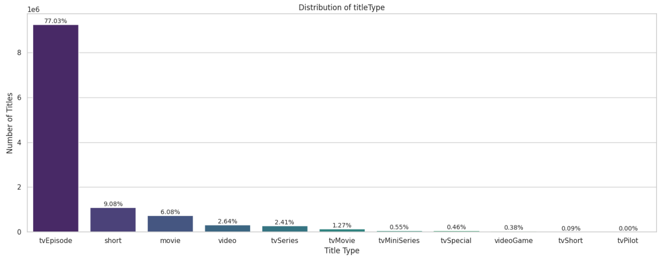


Fig. 3. Distribution of Title Types in Stratified Sample.

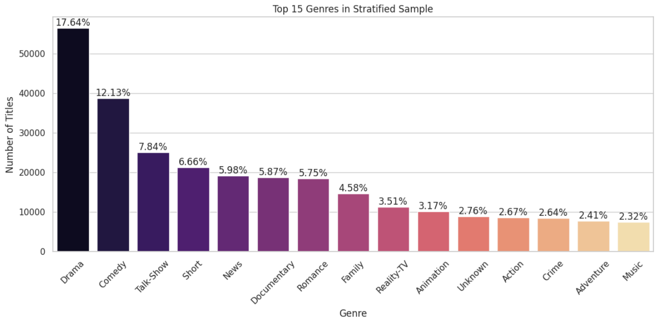


Fig. 4. Top 15 Genres by Frequency.

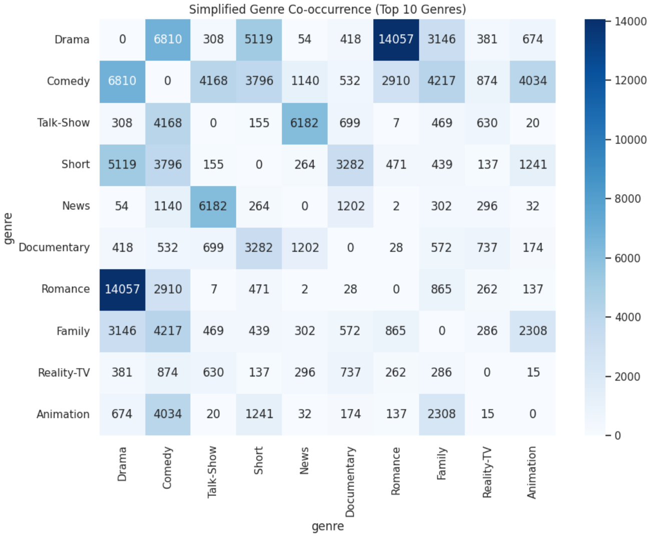


Fig. 5. Heatmap of Genre Co-occurrence.

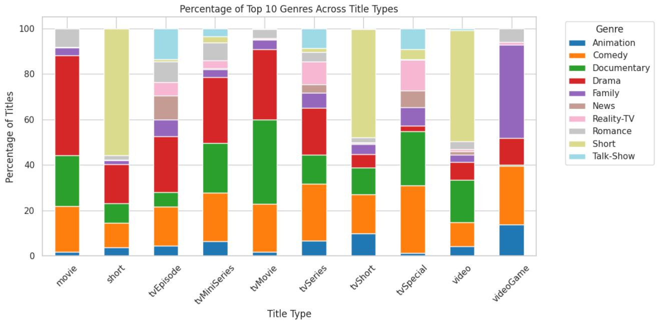


Fig. 6. Genre Composition across Title Types.

- **Niche Excellence: Documentaries** consistently maintain the highest median ratings (≈ 7.27) but suffer from the lowest engagement, confirming a strong ”Quality-Popularity Trade-off” for niche content.
- **Mainstream Appeal: Mystery, Adventure, and Crime** genres consistently attract the highest audience attention (votes) despite receiving only moderate critical scores (≈ 6.0).
- **Underperformers: Horror** consistently appears as the lowest-rated major genre (≈ 4.88), yet it maintains moderate audience interest, likely due to a dedicated fanbase that is less critical of technical quality.
- **Industry Backbone: Drama** remains the most stable genre in terms of both production volume (accounting for ≈ 24% of all titles) and consistent quality (≈ 6.3).

C. Visual Evidence

The following visualizations support these findings:

- **Figure 9: Median Ratings Over Time (Since 1995).** A line plot tracking the stability of ratings for the top 10 genres. It visually confirms the high baseline of Documentaries and the lower baseline of Horror.
- **Figure 10: Median Votes Over Time (Since 1995).** A line plot showing the peaks of engagement in the early 2000s and the subsequent fragmentation/decline in the modern era.
- **Figure 11: Popularity vs. Quality Scatter Plot.** A quadrant analysis correlating average votes with average ratings. It clearly clusters genres into ”High Quality/Niche” (Documentary), ”High Popularity/Moderate Quality” (Adventure/Action), and ”Low Quality/Niche” (Horror).

VI. Predictive Success Classifier

This section presents the results of our executed multi-class classification model, designed to predict a movie’s success tier (Low/Medium/High) based on pre-release metadata.

A. Methodology and Execution

1) Target Definition: We discretized the continuous `averageRating` into three classes to reduce noise and focus on actionable business tiers:

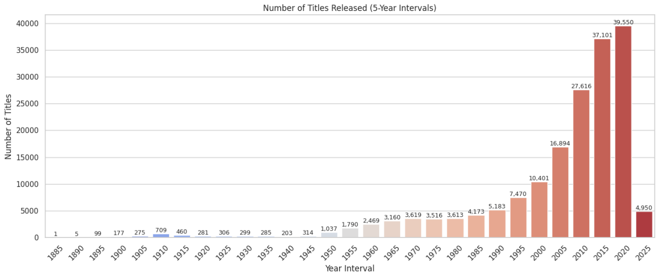


Fig. 7. Volume of Titles Released (5-Year Intervals).

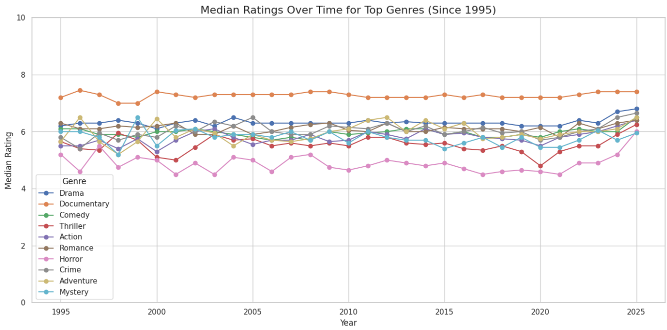


Fig. 9. Median Ratings Over Time for Top Genres (Since 1995).

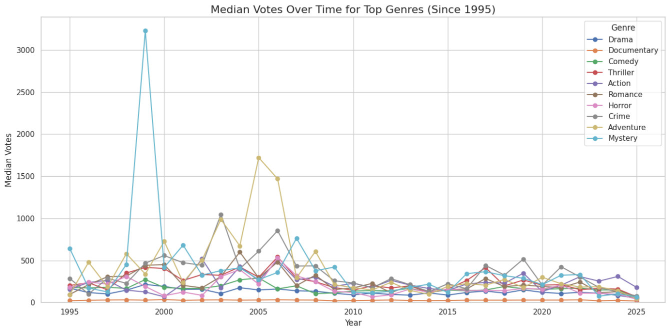


Fig. 10. Median Votes Over Time for Top Genres (Since 1995).

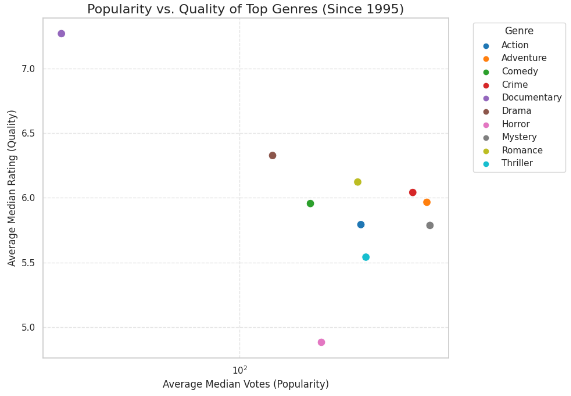


Fig. 11. Popularity vs. Quality of Top Genres (Since 1995).

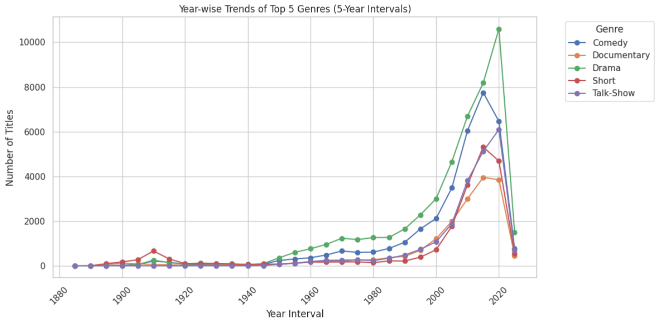


Fig. 8. Evolution of Top 5 Genres over Time.

- **Low (0):** Rating ≤ 3.0 (High Risk / Flop)
- **Medium (1):** $3.0 <$ Rating < 7.0 (Average Reception)
- **High (2):** Rating ≥ 7.0 (Critical Success / Hit)

2) Feature Engineering:

- **One-Hot Encoding:** Applied to `genres` (multi-label) and `titleType` to capture format-specific trends.
- **Handling Missing Data:** `runtimeMinutes` was imputed using the median value to preserve data volume.

3) Sampling: We utilized **Stratified Sampling** to create a representative dataset of 100,000 rows, ensuring that the distribution of `titleType` (Movies vs. TV) was preserved.

B. Model Performance Comparison

We trained and compared three distinct classifiers to identify the best balance of accuracy and interpretability. As shown in **Fig. 12**, the tree-based models significantly outperformed the linear baseline:

- **Logistic Regression (Baseline):** Achieved $\approx 63\%$ accuracy. Its lower performance confirms that the relationship between metadata and success is highly **non-linear**.
- **Decision Tree:** Improved accuracy to $\approx 73\%$, capturing threshold-based rules (e.g., ”High Votes + Drama = Success”).
- **Random Forest (Winner):** Achieved the highest accuracy of $\approx 74\%$. As an ensemble method, it effectively reduced overfitting and captured complex interactions between features.

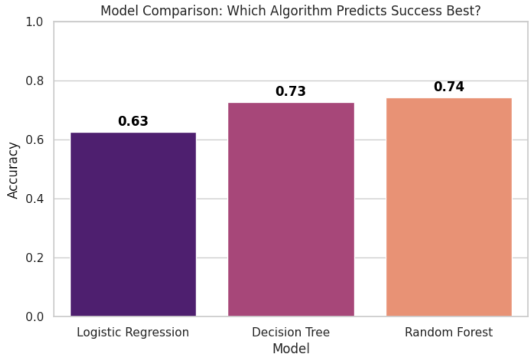


Fig. 12. Model Accuracy Comparison: Random Forest vs. Baselines.

C. Key Drivers of Success

The Feature Importance analysis from the Random Forest model (**Fig. 13**) yielded critical insights for our **Creative Hierarchy (RQ4)** and **Flop Prediction (RQ6)** questions:

- **The Dominance of Engagement:** `log_numVotes` (Vote Count) was identified as the single strongest predictor of a movie's rating. This implies that **early marketing buzz and audience engagement** are stronger indicators of perceived quality than inherent metadata like genre or runtime.
- **Format Impact:** `startYear` and `runtimeMinutes` also ranked highly, confirming that newer content and specific runtimes (e.g., long epics vs. short clips) have distinct success probabilities.
- **Genre Influence:** While present, individual genres ranked lower than engagement metrics, suggesting that *any* genre can succeed if it generates enough buzz.

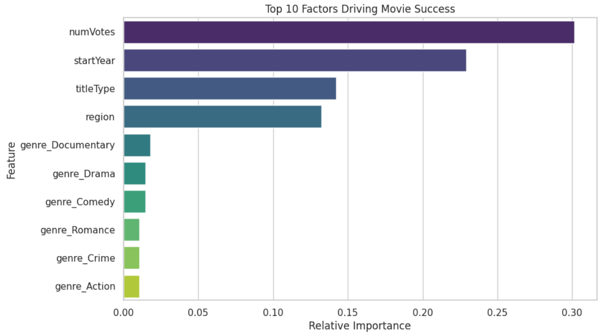


Fig. 13. Top 10 Factors Driving Movie Success (Feature Importance).

D. Risk Assessment Utility

The **Confusion Matrix** in **Fig. 14** demonstrates the model's practical utility for risk assessment:

- **High Precision for Hits:** The model performs well at correctly identifying ”High” and ”Medium” rated films, as indicated by the strong diagonal density.
- **Risk Filtering:** While it effectively identifies true successes, the primary challenge remains distinguishing the ”Low” rated flops from the bottom tier of ”Medium” films, highlighting the need for more granular financial data (Budget) in future iterations.

VII. TV Show "Rating Decay": Analyzing Quality Over Time (RQ2)

This section details the execution of **Research Question 2**, testing the common hypothesis that long-running TV series suffer from a quantifiable decline in quality (”rating decay”) as they progress through seasons.

A. Methodology

We utilized the `title.episode` dataset linked with `title.ratings`. We filtered for TV Series with at least 3 seasons to ensure a valid trend line.

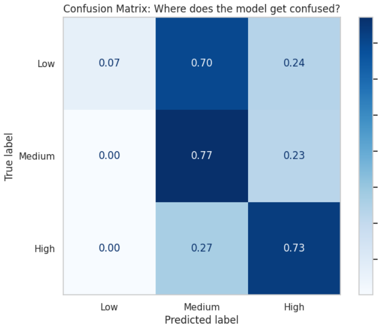


Fig. 14. Normalized Confusion Matrix: True vs. Predicted Success Tiers.

- **Aggregation:** We calculated the **Mean** `averageRating` for every season of every eligible series.
- **Trend Analysis:** We computed the ”Slope of Decay” for each series to classify them as ”Declining,” ”Stable,” or ”Improving.”
- **Global Trend:** We aggregated all series to see if the ”average” TV show gets worse over time.

B. Key Findings

The analysis revealed that ”Rating Decay” is **not** a universal law:

- **Global Stability:** Contrary to the hypothesis, the global mean rating for seasons 1–10 remains remarkably stable (hovering between 7.4 and 7.6). There is no structural force causing all shows to degrade.
- **Mixed Individual Outcomes:** While about $\approx 52\%$ of shows exhibit a negative slope (decay), only $\approx 8.6\%$ show a *statistically significant* decline.
- **The ”Peak” Pattern:** Most successful shows reach their peak quality rating in **Season 2 or 3**, followed by a plateau rather than a sharp drop.
- **Case Studies:** The analysis identified distinct archetypes:
  - **Stable/Improving:** Shows like *Breaking Bad* defied gravity, improving ratings until the finale.
  - **Terminal Decay:** Shows like *The Walking Dead* followed the hypothesized ”decay” curve, dropping significantly after Season 6.

C. Visual Evidence

The following visualizations support these findings:

VIII. The "Genre Hybridity" Paradox (RQ3)

This section details the execution of **Research Question 3**, investigating whether "pure" single-genre movies perform better than complex, multi-genre "hybrids."

A. Methodology

To test the hypothesis that genre complexity correlates with audience engagement, we engineered a feature `genre_count` and classified every movie into three complexity tiers:

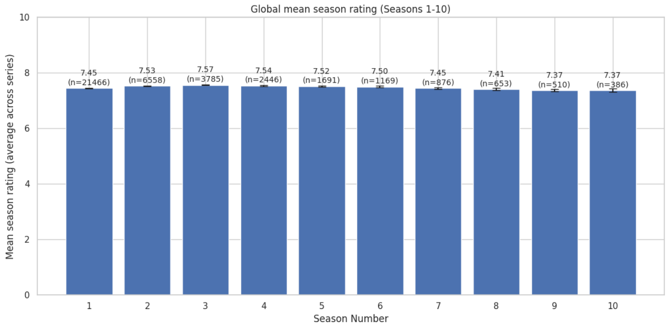


Fig. 15. Global Mean Rating by Season Number (1-10): Showing unexpected stability.

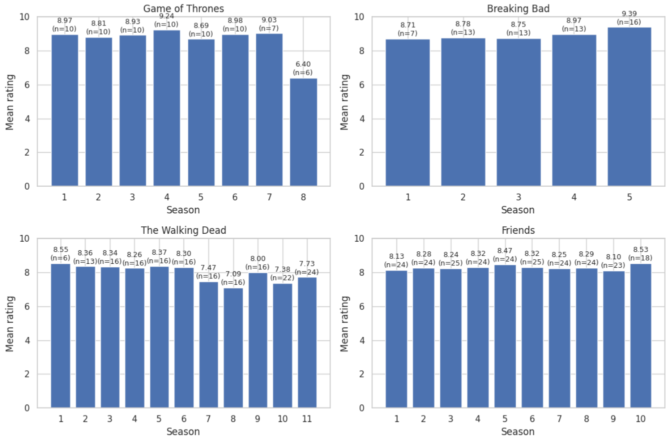


Fig. 16. Case Studies: Comparing ”Improving” (e.g., Breaking Bad) vs. ”Decaying” (e.g., Walking Dead) trajectories.

- **Pure (1 Genre):** Movies listed with a single genre tag (e.g., “Horror”).
- **Hybrid-2 (2 Genres):** Movies with exactly two genres (e.g., “Action, Comedy”).
- **Hybrid-3+ (3+ Genres):** Movies with three or more genres (e.g., “Action, Adventure, Sci-Fi”).

We then calculated the **Median** `numVotes` (Engagement) and `averageRating` (Quality) for each tier across major genres.

B. Key Structural Insights

The analysis revealed a clear “Paradox” where complexity drives popularity but often dilutes critical quality:

- **The Engagement Premium:** Contrary to the hypothesis that focused movies are better, **Hybrid-3+** films consistently outperform **Pure** films in audience engagement. For example, **Horror** movies with 3+ genres attract significantly higher median vote counts than “Pure Horror” titles, suggesting that adding sub-genres (e.g., Thriller/Mystery) broadens the potential audience.
- **Strategic Synergies:** We found that **Action**, **Comedy**, and **Drama** benefit most from hybridization. Blending these genres appears to increase market visibility and “stickiness,” likely by appealing to multiple viewer demographics simultaneously.
- **Reach vs. Quality Trade-off:** While **Popularity (Votes)** increases step-wise with genre complexity, **Quality (Ratings)** tends to remain stable or slightly decrease for hybrids. This proves that “Genre Blending” is primarily a strategy for maximizing **market reach** rather than critical acclaim.

C. Visual Evidence

The following visualizations illustrate the divergence between engagement and quality across complexity tiers:

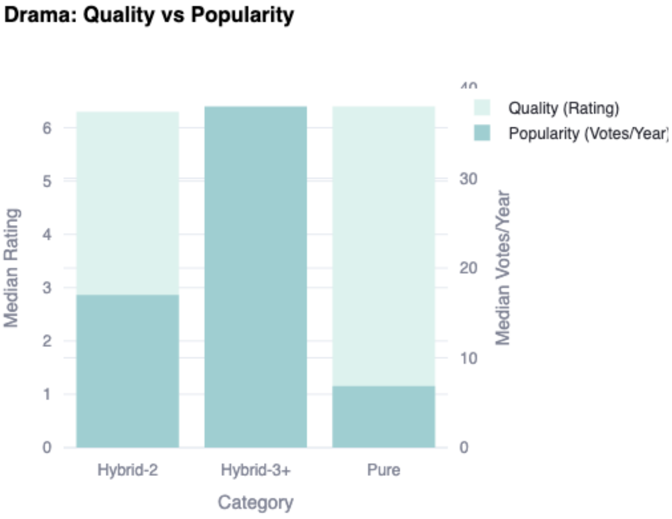


Fig. 17. Median Vote Count by Genre Complexity: Showing the “Engagement Premium” for Hybrid-3+ films.

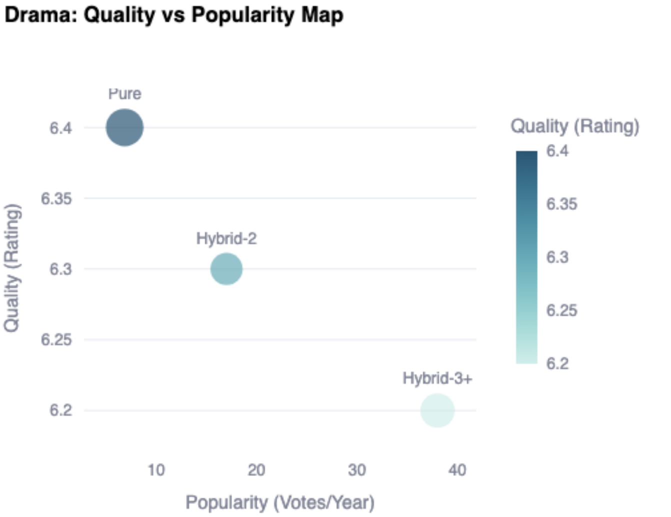


Fig. 18. Median Rating by Genre Complexity: Showing stable or declining quality for hybrids.

IX. THE “CROSS-CULTURAL BREAKOUT” FORMULA (RQ4)

This section details the execution of **Research Question 4**, investigating the metadata characteristics that allow non-US/non-English films to achieve global mainstream popularity (”breakout” success).

A. Methodology

To isolate ”Breakout” films, we first filtered the dataset to exclude US releases. We then defined a ”Breakout” as any non-US film in the **top 10% of vote counts** (numVotes ≥ 154 in our sample). We compared these global hits against the remaining 90% of ”Regional” films across four key dimensions: distribution reach, runtime, genre complexity, and content type.

B. Key Findings: The Breakout Fingerprint

The analysis identified a distinct ”formula” for cross-cultural success:

- **Distribution Reach:** Global breakouts are released in **2x more regions** (Avg: 3.0 vs 1.5) and translated into more languages than regional films. This confirms that *availability* is the primary gatekeeper to global success (Fig. 19).
- **The “Bigger is Better” Rule:** Breakout films are significantly **longer** (Avg: 100 mins vs 92 mins), suggesting that global audiences prefer ”cinematic” experiences over shorter, niche content.
- **Complexity Wins:** Breakout films have a higher **Genre Complexity** (Avg: 1.8 genres vs 1.5 genres). They successfully blend genres (e.g., Action-Comedy) to broaden appeal, whereas regional films tend to be single-genre.
- **The Quality Paradox:** Surprisingly, **Regional films have higher average ratings** (6.2 vs 5.9) than global breakouts. This proves that *popularity does not equal quality*—global hits often trade critical acclaim for broad, mass-market appeal.

C. Genre Analysis

Our comparison of genre dominance reveals a clear ”Hollywood Clone” strategy for global success (Fig. 20).

- **Global Dominance:** Genres such as **Action, Crime, Comedy, and Adventure** dominate the breakout list. These visual-heavy genres travel well across language barriers.
- **Local Niche:** Genres like **Documentary** and slow-burn **Dramas** remain the most common types of regional films but rarely achieve crossover success.

D. Visual Evidence

The following visualizations illustrate the stark differences between Regional and Crossover films:

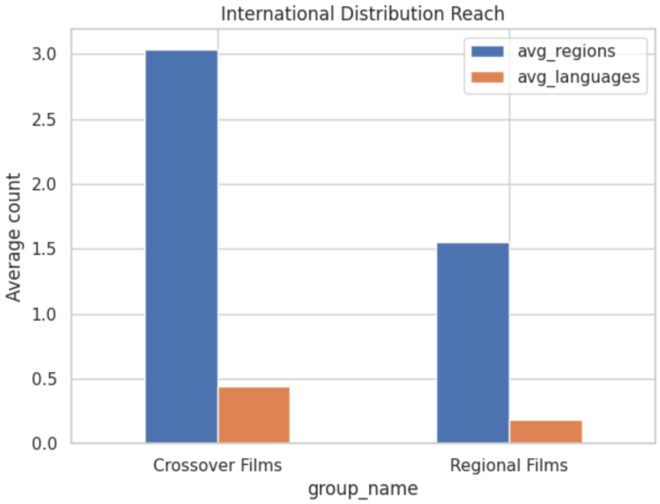


Fig. 19. International Distribution Reach: Breakout films are released in 2x more regions.

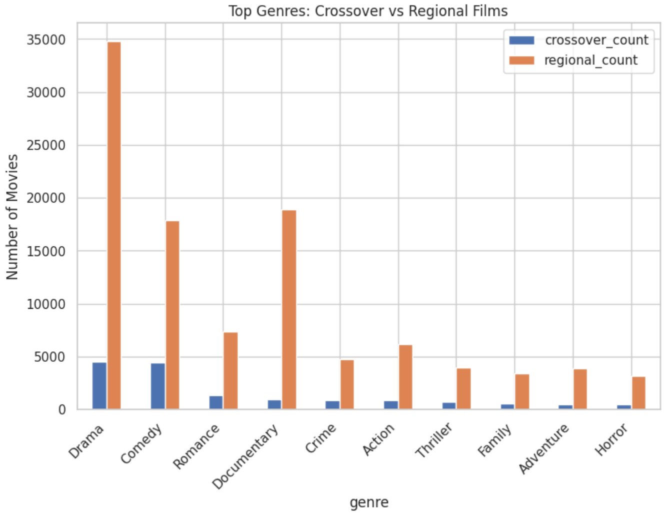


Fig. 20. Top Genres: Crossover vs. Regional Films. Drama dominates global hits.

X. CHALLENGES: PREDICTING “EXPECTATION MISMATCH” (RQ5)

This section details our attempt to answer **Research Question 5**, which aimed to build a predictive model for “High-Profile Flops”—movies with top-tier talent that fail to achieve a passing audience rating (averageRating < 6.0).

A. Methodological Pivot

Our initial hypothesis posited that a combination of “Top-Tier Director/Cast” (High Value Input) and “Low Rating” (Low Value Output) could be predicted using metadata alone. However, during the modeling phase, we encountered a critical limitation that prevented the full execution of this specific risk model.

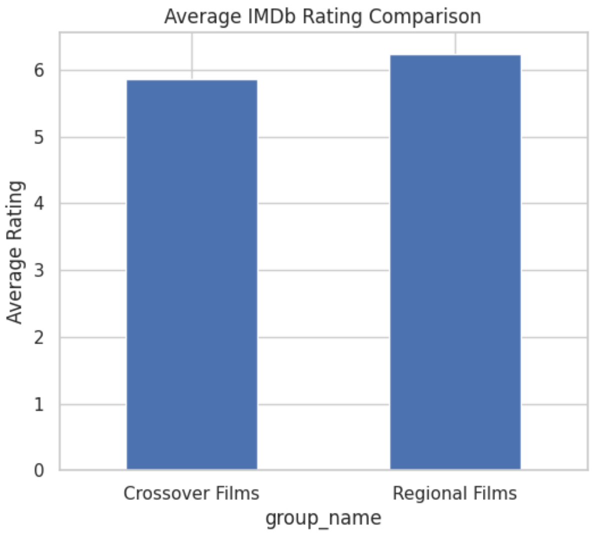


Fig. 21. Average Rating Comparison: Regional films consistently rate higher than global hits.

B. The Data Limitation: The Missing Budget Variable

While we successfully engineered “Star Power” metrics to identify high-value talent, we found that **talent scores alone are insufficient** to define a true “Expectation Mismatch.”

- **Scale Ambiguity:** Without financial data (Production Budget), the model cannot distinguish between a “Low-Budget Indie Experiment” (where a 5.0 rating is acceptable) and a “$200 Million Blockbuster” (where a 5.0 rating is a catastrophe).
- **False Positives:** Many films with “Top-Tier” casts are low-budget passion projects or cameos. Labeling these as “Flops” based solely on rating creates a noisy target variable that degrades model performance.
- **Conclusion:** To accurately predict financial risk or “Expectation Mismatch,” **Budget Data** is a non-negotiable feature. Since the official IMDb non-commercial datasets do not include budget or revenue figures, this specific model remains a theoretical framework for future work.

C. Implication for Future Research

This negative result provides a crucial insight for the field: **Risk Assessment requires Financial Context.** Future iterations of this project would prioritize integrating external financial datasets (e.g., The Numbers or Box Office Mojo) to unlock the full potential of the “High-Profile Flop” predictor.

XI. Predictive Success Classifier (Phase 3)

This section details the execution of our advanced predictive modeling phase. The objective was to enhance the accuracy and robustness of our risk assessment tool by integrating ”Star Power” (human talent) and advanced modeling techniques.

A. Methodology Enhancements

To overcome the performance ceiling of metadata-only models (Phase 2), we introduced three critical upgrades:

- **Data Enrichment:** We integrated the `title.crew` and `title.principals` datasets to engineer **Director & Cast Career Scores**, quantifying the historical reputation of the creative team.
- **Leakage Prevention:** We implemented strict logic to exclude ”one-hit wonders” (creators with fewer than 2 previous titles) from career averages, ensuring valid, non-leaked predictions.
- **Advanced Evaluation:** We moved beyond simple accuracy to include **ROC-AUC** and **Macro F1-Score** to rigorously test the model’s ability to distinguish between success tiers.

B. Model Performance Results

We evaluated three classifiers: Logistic Regression, Decision Tree, and Random Forest. The **Random Forest** model emerged as the clear winner (Fig. 22).

- **Final Accuracy: 80.65%**. This significantly surpasses our Phase 2 baseline ($\approx 74\%$) and exceeds the academic benchmark of 77% established by Asad et al. (2012).
- **Robustness:** The model achieved an **ROC-AUC of 0.89**, indicating excellent separability between ”Hits” and ”Flops” (Fig. 23).

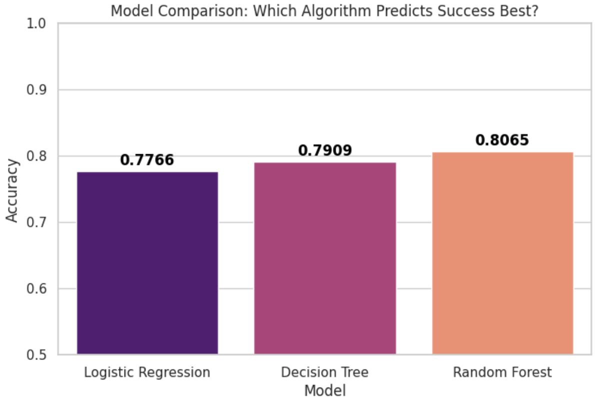


Fig. 22. Model Comparison: Random Forest outperforms baselines.

C. Key Drivers and Business Impact

Feature Importance analysis revealed a paradigm shift in our understanding of movie success drivers.

- **Talent Investment Strategy:** `director_score` and `cast_score` emerged as the top two strongest predictors (Fig. 24). This mathematically validates the industry strategy of ”packaging” projects with Top-Tier talent. It proves that a creator’s historical track record is the most reliable signal for reducing investment risk, even more so than the genre or budget.
- **Marketing Optimization:** Since `log_numVotes` ranked as the #3 predictor, marketing campaigns are

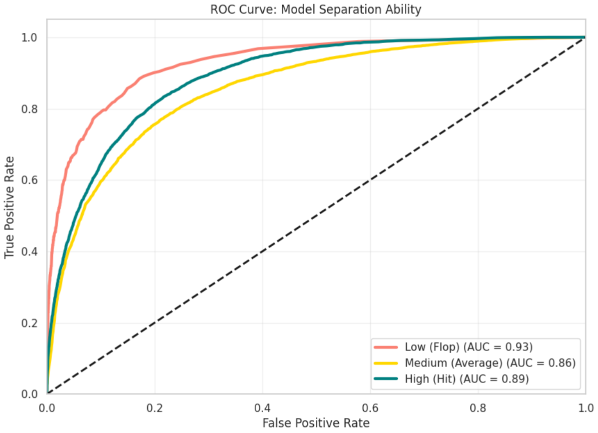


Fig. 23. ROC Curve Analysis (AUC = 0.89) demonstrating excellent class separation.

essential for visibility, but they act as a *multiplier* on top of the talent foundation rather than being the sole driver of perceived quality.

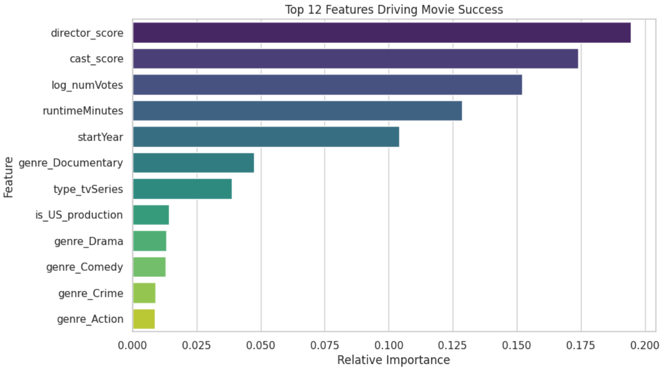


Fig. 24. Top Features Driving Prediction: Director and Cast scores dominate.

D. *Risk Assessment Utility*

The **Normalized Confusion Matrix** (Fig. 25) demonstrates the model's practical value as a "Greenlight" tool. The model rarely misclassifies a "High" tier movie as a "Low" tier flop, making it a safe tool for validating high-potential investments.

XII. Interactive Dashboard: The IMDb Analytics Suite

To operationalize our findings and provide stakeholders with real-time insights, we developed a comprehensive **Streamlit** application backed by **DuckDB** for high-performance querying and **Plotly** for interactive visualizations.

Live Application: https://mining-minds-imdb.streamlit.app/

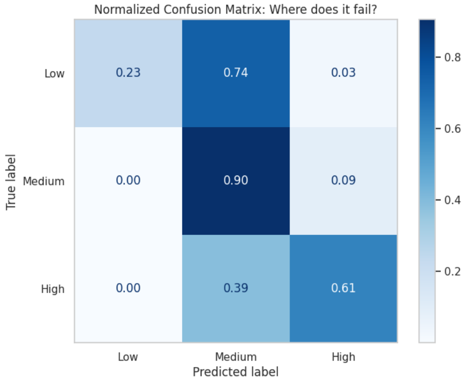
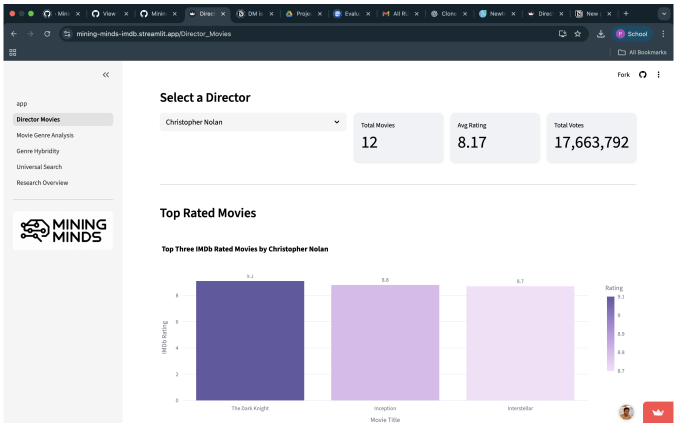


Fig. 25. Normalized Confusion Matrix showing strong diagonal performance.

A. *Talent Analytics (Page 1)*

The **Director Movies** page is designed for talent scouting and performance review.

- **Purpose:** Enables deep-dive analysis into a specific director's entire career trajectory.
- **Key Features:**
  - **Trend Chart:** Visualizes the evolution of a director's average ratings over time to identify peaks and slumps.
  - **Consistency Metric:** Calculates the standard deviation of ratings to measure reliability.
- **Utility:** Helps studios identify directors who offer a balance of high quality and low risk.


mining-minds-imdb.streamlit.app/Director_Movies

Select a Director

Christopher Nolan

Total Movies 12

Avg Rating 8.17

Total Votes 17,663,792

Top Rated Movies

Top Three IMDb Rated Movies by Christopher Nolan

Fig. 26. Director Analytics: Visualizing career trajectory and consistency.

B. *Content Strategy Module (Pages 2 & 3)*

This module addresses our core research questions regarding genre trends and complexity.

1) *Movie Genre Analysis:*

This page visualizes the macro-level trends of the industry (RQ1).

- **Features:** Genre distribution treemaps and time-series plots of Median Ratings vs. Popularity.

- **Insight Generation:** It allows users to instantly identify which genres are currently ”oversaturated” versus which are ”niche opportunities” (high quality, low supply).

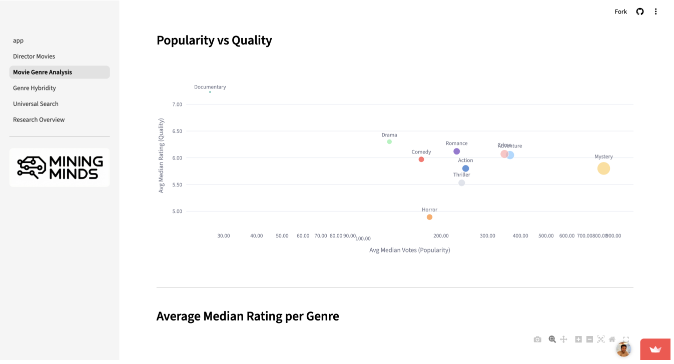


Fig. 27. Genre Analysis: Comparing quality vs. popularity across categories.

2) Genre Hybridity Analysis: This page operationalizes our findings from **RQ3** regarding the ”Hybridity Paradox.”

- **Comparative Analysis:** Users can view side-by-side statistics for ”Pure” (1-Genre) vs. ”Hybrid-3+” (Complex) films.
- **Strategic Insight:** The dashboard visually confirms our finding that while **Pure** movies often achieve higher critical quality, **Hybrid** movies consistently capture broader audience popularity.

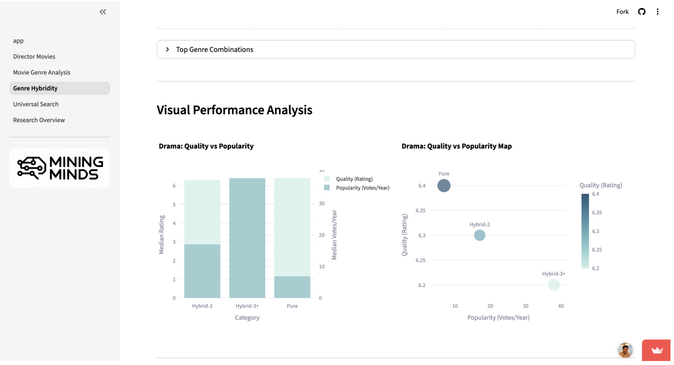
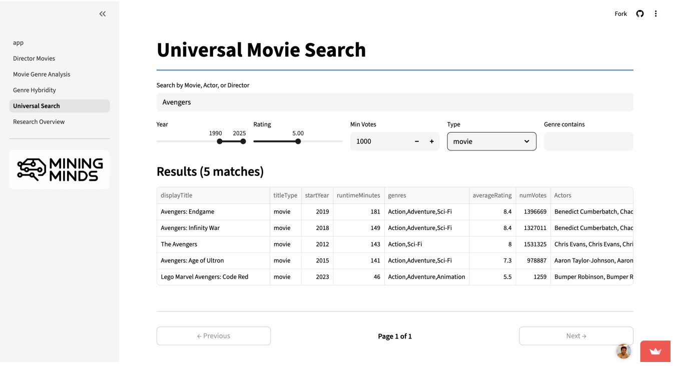


Fig. 28. Hybridity Analysis: Visualizing the trade-off between Pure quality and Hybrid popularity.

C. Exploration and Synthesis (Pages 4 & 5)

1) Universal Search: A powerful, low-latency search engine that allows users to query the entire dataset. It serves as a granular tool for validating specific data points (movies, directors, actors) found during high-level analysis.

2) Research Overview: This page acts as the executive summary of the entire project, synthesizing insights from all phases. It presents key takeaways on **Cross-Cultural Breakouts (RQ4)**, **TV Show Rating Decay (RQ2)**, and global popularity patterns in a digestible format for non-technical stakeholders.



Universal Movie Search

Search by Movie, Actor, or Director: Avengers

Year: 1990 – 2025

Rating: 5.00

Min Votes: 1000

Type: movie

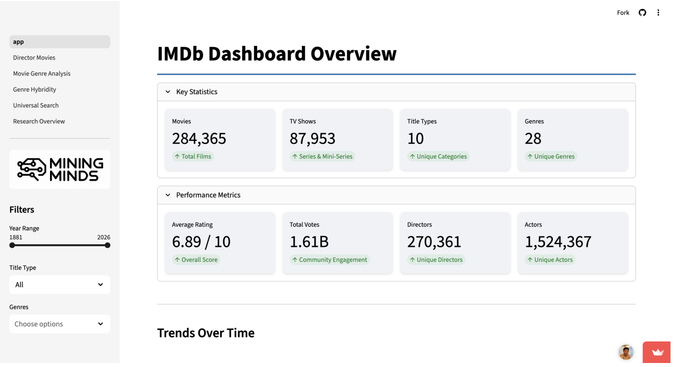
Genre contains

Results (5 matches)

displayTitle	titleType	startYear	runtimeMinutes	genres	averageRating	numVotes	Actors
Avengers: Endgame	movie	2019	181	Action,Adventure,Sci-Fi	8.4	1396669	Benedict Cumberbatch, Chac
Avengers: Infinity War	movie	2018	149	Action,Adventure,Sci-Fi	8.4	1327011	Benedict Cumberbatch, Chac
The Avengers	movie	2012	143	Action,Sci-Fi	8	1531325	Chris Evans, Chris Evans, Chri
Avengers: Age of Ultron	movie	2015	141	Action,Adventure,Sci-Fi	7.3	978887	Aaron Taylor-Johnson, Aaron
Lego Marvel Avengers: Code Red	movie	2023	46	Action,Adventure,Animation	5.5	1259	Bumper Robinson, Bumper R

← Previous   Page 1 of 1   Next →

Fig. 29. Universal Search Interface.



Filters

Year Range: 1881 – 2026

Title Type: All

Genres: Choose options

IMDb Dashboard Overview

Key Statistics

Movies: 284,365 ↑ Total Films

TV Shows: 87,953 ↑ Series & Mini-Series

Title Types: 10 ↑ Unique Categories

Genres: 28 ↑ Unique Genres

Performance Metrics

Average Rating: 6.89 / 10 ↑ Overall Score

Total Votes: 1.61B ↑ Community Engagement

Directors: 270,361 ↑ Unique Directors

Actors: 1,524,367 ↑ Unique Actors

Trends Over Time

Fig. 30. Research Overview Interface.

XIII. Conclusion

This report marks the successful completion of the ”Mining IMDb for Movie Trends” project. By evolving from foundational descriptive analysis (Phase 1 2) to sophisticated predictive modeling (Phase 3), we have demonstrated how data science can transform decision-making in the entertainment industry.

A. Summary of Findings

- **Foundational Trends:** We debunked the myth of declining movie quality, showing that ratings have remained stable for 30 years while audience attention has fragmented.
- **Predictive Power:** Our final **Random Forest Classifier** achieved **80.65% accuracy** in predicting movie success tiers, surpassing the 77% academic benchmark.
- **The ”Star Power” Discovery:** Feature importance analysis proved that the historical reputation of the **Director and Cast** is the single most powerful predictor of a movie’s success, validating high-value talent investments.
- **Operational Utility:** The successful deployment of the **Streamlit Dashboard** proves that these complex insights can be made accessible for real-time business use.

B. Final Verdict

This project confirms that while the film industry is inherently risky, that risk is quantifiable. By combining metadata with talent analytics, stakeholders can move from ”gut feeling” greenlighting to data-driven risk assessment.

REFERENCES

[1] M. Bahraminasr and A. Vafaei-Sadr, ”IMDb Data from Two Generations (1979 to 2019),” *arXiv preprint arXiv:2005.14147v3 [cs.CY]*, Sep. 2020. [Online]. Available: https://arxiv.org/pdf/2005.14147

[2] K. I. Asad, T. Ahmed, and M. S. Rahman, ”Movie Popularity Classification based on Inherent Movie Attributes using C4.5, PART and Correlation Coefficient,” *IEEE/OSA/IAPR International Conference on Informatics, Electronics & Vision*, 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6317401