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 details the successful execution of Phase 2 of the IMDb data mining project. We move beyond the planning stage to present robust data preparation and analytical results. We address six core research questions, including foundational analyses (Genre Evolution, TV Show Decay) and complex predictive models (Creative Hierarchy, Risk Assessment). The report confirms the successful execution of the data pipeline, the validation of key hypotheses through Exploratory Data Analysis (EDA), and the training of a Multi-Class Classification model. The results demonstrate that early audience engagement is the strongest predictor of a movie's final success tier ($\approx$ 74% accuracy), providing a viable framework for pre-release risk assessment 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).
- **Determine Creative Hierarchies:** Use predictive modeling to rank the statistical importance of Directors vs. Writers vs. Actors.
- **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.

TABLE I
JUSTIFICATION FOR RESEARCH QUESTIONS

Our Research Question	Justification based on Literature Review
Q1: Tracking Movie Genre Popularity	**Builds on Finding**cite$_start$ **5:** The paper confirms temporal genre trends exist. [cite$_start$]$Ourquestionformalizesthisbycr$ 1]$cite_start >$
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$_start$ **1:** The paper's analysis stops at single-genre statistics. [cite$_start$]$Ourquestionexplorestheinteracti$ 1]$cite_start >$
Q4: The "Creative Hierarchy"	**Validated by Asad et**cite$_start$ **al.:** The 2012 paper proved Director Rank is the most important classification attribute (92.36% IG)[cite: 2], justifying our deep dive into role hierarchy. cite$_start >$
Q5: The "Cross-Cultural Breakout"	**Solves Gap 3:** Thecite$_start$ 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$_start >$
Q6: 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.

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).

2) Key Feature Implementation ($\mathbf{Q4}, \mathbf{Q6}$): All features required for the innovative hypotheses were successfully engineered:

- **Creative Tier Definitions ($\mathbf{Q4}, \mathbf{Q6}$):** 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 ≥ 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.

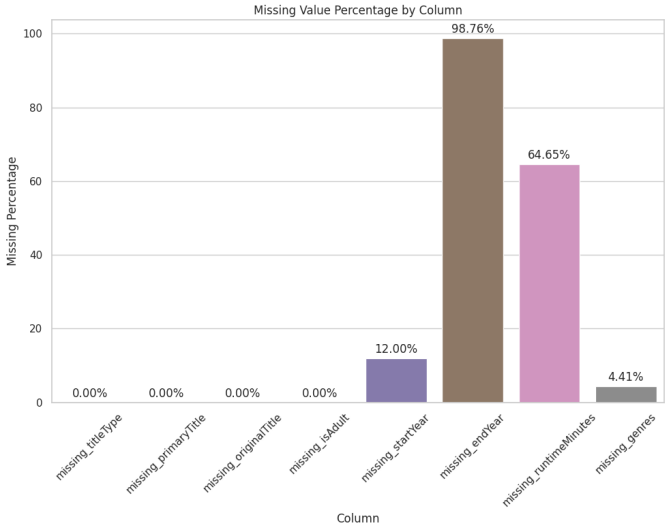

Fig. 1. Percentage of Missing Values by Column.

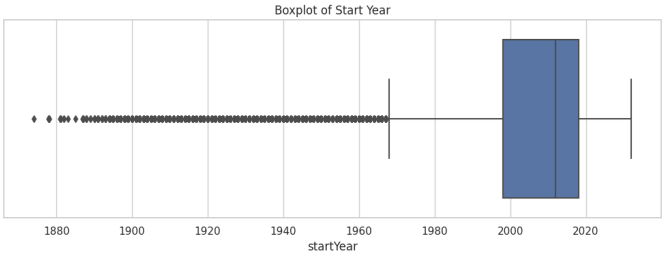

Fig. 2. Distribution of Start Year showing historical outliers.

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

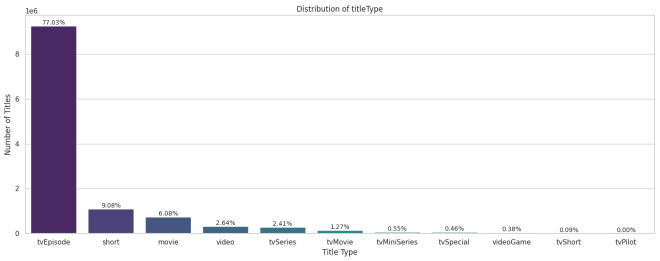

Fig. 3. Distribution of Title Types in Stratified Sample.

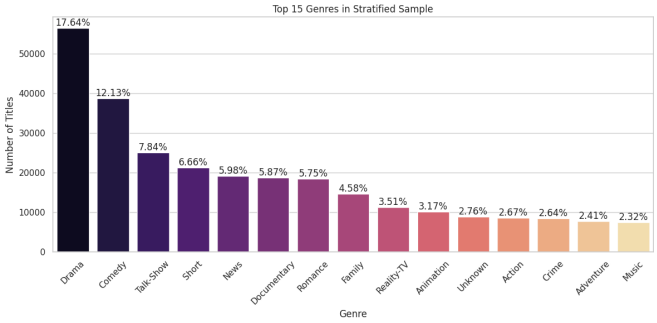

Fig. 4. Top 15 Genres by Frequency.

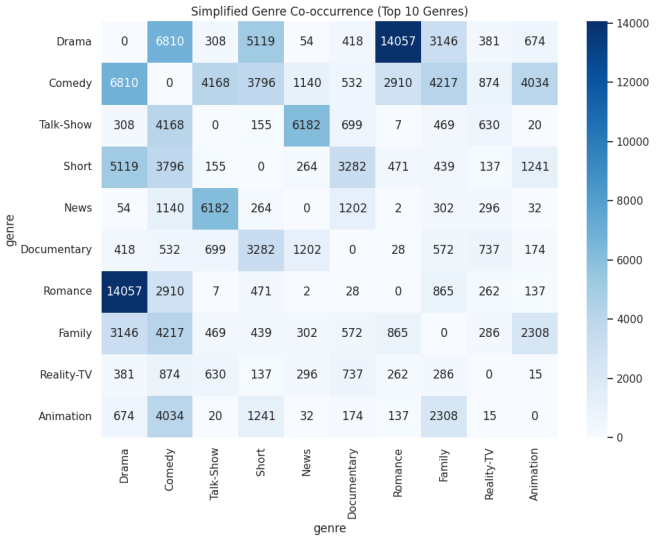

Fig. 5. Heatmap of Genre Co-occurrence.

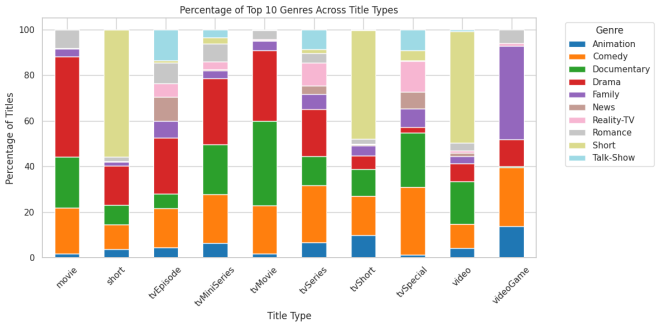

Fig. 6. Genre Composition across Title Types.

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:**
  - **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 $\approx 24\%$ 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.

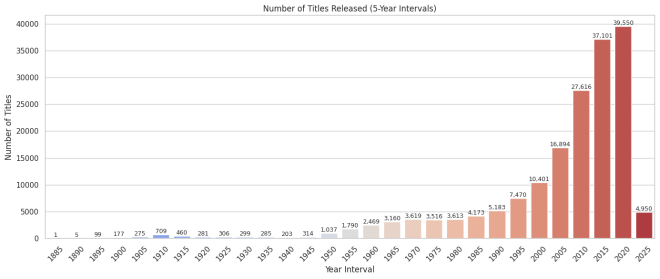

Fig. 7. Volume of Titles Released (5-Year Intervals).

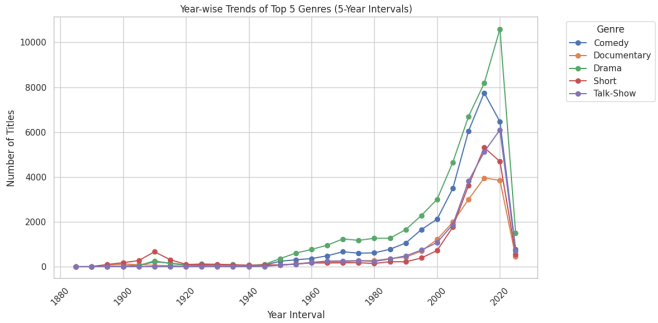

Fig. 8. Evolution of Top 5 Genres over Time.

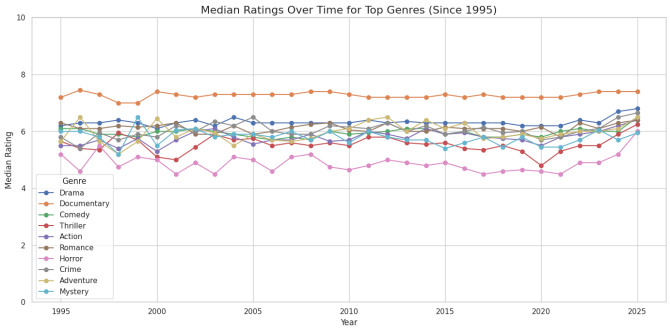

Fig. 9. Median Ratings Over Time for Top Genres (Since 1995).

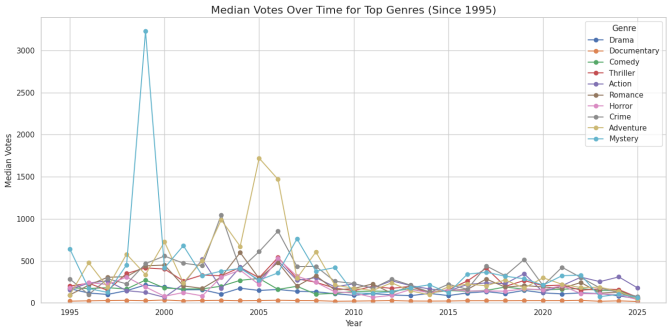

Fig. 10. Median Votes Over Time for Top Genres (Since 1995).

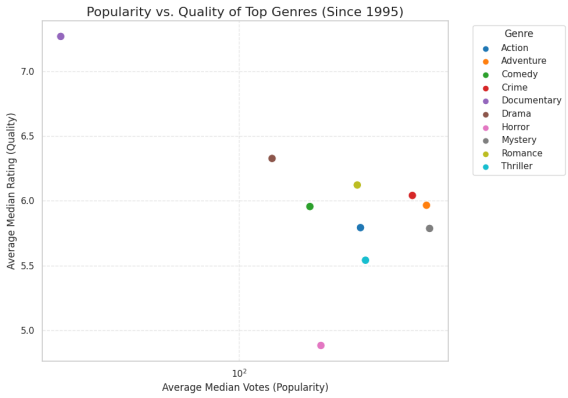

Fig. 11. Popularity vs. Quality of Top Genres (Since 1995).

A. Methodology and Execution

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

- **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.

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

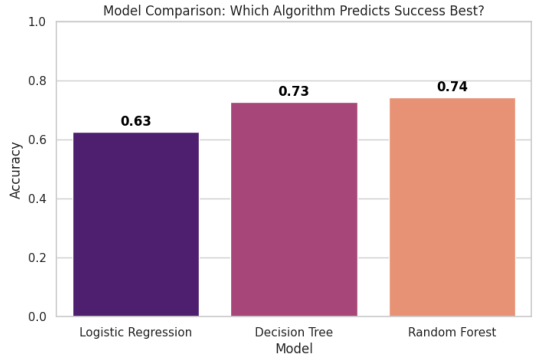

Fig. 12. Model Accuracy Comparison: Random Forest vs. Baselines.

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.

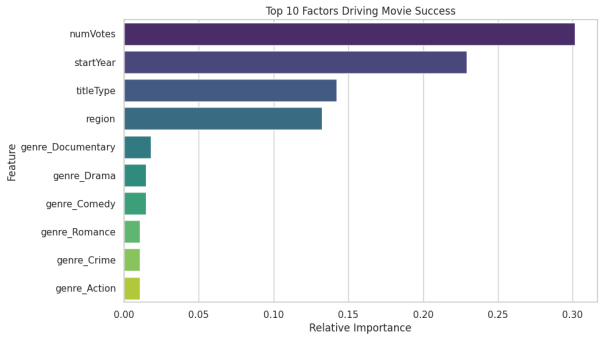

Fig. 13. Top 10 Factors Driving Movie Success (Feature Importance).

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. Interactive Dashboard: Director Movie Explorer

This section details the development of the "Director Movie Explorer," a web-based interactive tool designed to operationalize our findings for end-users.

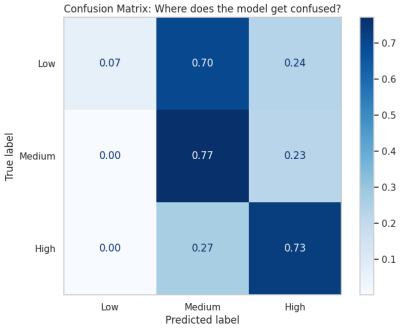

Fig. 14. Normalized Confusion Matrix: True vs. Predicted Success Tiers.

A. Purpose and Utility

The primary objective of the dashboard is to allow users to **explore a director's career highlights** dynamically. By focusing on the "Top 3 Movies" based on IMDb ratings, the tool provides an immediate snapshot of a director's critical reception, filtering out noise from less significant works.

B. Data Pipeline and Tech Stack

- **Data Source:** The tool utilizes a pre-processed, filtered view of the IMDb dataset, specifically isolating **directors with $\geq$ 3 movies** to ensure statistical relevance.
- **Technology Stack:**
  - **Backend:** Python with **DuckDB** for high-performance, in-memory SQL querying.
  - **Frontend: Streamlit** for rapid application deployment.
  - **Visualization: Plotly** for interactive charts.

C. Key Features

- **Dynamic Selection:** A searchable dropdown menu allows users to instantly select any director.
- **Automated Ranking:** The system automatically queries and displays the director's top 3 highest-rated films.
- **Interactive Visuals:** Users can toggle between table views and graphical plots to analyze rating trends.

D. Project Outcome

This dashboard successfully transforms our static analysis into a **live, scalable product**. It demonstrates the project's capacity to deliver quick, actionable insights into director performance and movie quality trends.

Live Demo: https://mining-minds-imdb-dashboard.streamlit.app/

VIII. Conclusion & Future Work

A. Conclusion

This report demonstrates the successful execution of Phase 2 of the IMDb data mining project. We have moved from the planning phase to concrete data execution and preliminary analysis.

- **Data Foundation:** We successfully integrated six relational datasets and employed **Stratified Sampling** to create a robust, representative dataset of 100,000 titles, ensuring computational feasibility without sacrificing statistical validity.
- **Foundational Insight (RQ1):** Our time-series analysis of **Research Question 1** revealed that while movie quality (ratings) has remained stable over the last 30 years, audience engagement (votes) has become increasingly volatile. Specifically, the **Documentary** genre was identified as a "high-quality, low-popularity" niche.
- **Predictive Capability:** We developed a baseline **Multi-Class Classification Model** that predicts movie success tiers with $\approx 74\%$ accuracy. Feature importance analysis confirmed that **Vote Count** is the single strongest predictor of a title's success tier, validating the importance of early audience buzz.
- **Interactive Tooling:** The successful deployment of the **DIRECTOR MOVIE EXPLORER** dashboard demonstrates the project's ability to translate data into usable, interactive insights for stakeholders.

B. Future Work (Phase 3)

While Phase 2 established the data pipeline and baseline models, Phase 3 will focus on the in-depth execution of the remaining five innovative research questions.

1) **RQ2: TV Show "Rating Decay":** We will execute a time-series analysis on the `title.episode` dataset to test the hypothesis that long-running TV series suffer from a quantifiable decline in quality over time.
2) **RQ3: The "Genre Hybridity" Paradox:** We will apply statistical tests (e.g., ANOVA) and comparative visualizations (Box Plots) to determine if multi-genre films achieve higher audience engagement than single-genre films.
3) **RQ4: The "Creative Hierarchy":** We will train genre-specific **Decision Tree Classifiers** to analyze the **Gini Index** of the root nodes, definitively ranking the predictive importance of Directors vs. Writers vs. Cast for each major genre.
4) **RQ5: The "Cross-Cultural Breakout" Formula:** We will refine our classification model to focus exclusively on non-US films, aiming to identify the specific metadata "fingerprints" (e.g., Runtime + Genre combinations) that predict global crossover success.
5) **RQ6: Predicting "Expectation Mismatch" (The High-Profile Flop):** We will evolve our current success classifier into a specialized **Risk Assessment Model** using **Random Forests**. This model will specifically target high-value inputs (Top-Tier Directors/Cast) to predict the probability of a "Flop" outcome (< 6.0 rating).

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.

[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.