Mining IMDb for Movie Trends: Phase one, Dataset Introduction and Preliminary Analysis

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

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

***Abstract*—This paper outlines the foundational methodology for a data mining project aimed at uncovering high-value, novel patterns from the IMDb (Internet Movie Database) dataset. Moving beyond standard descriptive analysis, our research focuses on complex, second-order effects. We introduce six research questions: two foundational analyses (movie genre trends, TV show rating decay) and four innovative models ("Genre Hybridity" paradox, "Creative Hierarchy" of talent, "Cross-Cultural Breakout" formula, and "High-Profile Flop" risk assessment). This paper details our literature review, which identifies clear research gaps, our comprehensive, syllabus-compliant data pre-processing pipeline, and the specific analytical models (e.g., Decision Trees, Random Forests, Aggregation) required to test each hypothesis. This work provides a complete plan for generating novel, actionable insights into the drivers of film and television success.**

***Index Terms*—Data Mining, IMDb, Machine Learning, Research Methodology, Data Pre-processing, Classification, Feature Engineering, Syllabus-Compliant**

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."

C. Paper Structure

This paper details the complete plan for this project. Section II reviews the academic literature and identifies our research gaps. Section III presents our six core research questions and their corresponding hypotheses. Section IV details the data pre-processing pipeline. Finally, Section V concludes the paper.

II. Literature Review

Our research is academically grounded and designed to fill specific gaps left by existing literature.

A. Overview of Foundational Research

Our review is centered on the foundational paper "IMDb Data from Two Generations (1979 to 2019)" by Bahraminasr and Vafaei-Sadr (2020) [1]. This article presents a comprehensive, custom-scraped IMDb-based dataset. Using statistical analysis, the study reveals important trends in film ratings, vote distributions, and demographic influences on audience reception.

B. Source Reliability and Authenticity

This paper is considered a reliable academic source. It was published on arXiv (arXiv:2005.14147v3), a standard, highly-respected open-access repository for preprints in computer science [1]. Its public availability and clear methodology make it an appropriate foundation for our review.

C. Key Findings from the Paper

The authors' analysis revealed several key findings that validate our project's core assumptions.

- **Finding 1: Metadata Patterns are Measurable:** The study confirms that metadata has a measurable relationship with ratings, finding clear differences in average ratings by MPAA certificate and Genre [1].
- **Finding 2: Temporal Trends are Significant:** The paper proves that movie production (FIG. 1), vote counts (FIG.

2), and average ratings (FIG. 4) all show clear trends over time [1].

- **Finding 3: Demographic & Regional Bias is Proven:** A core finding is the demonstrable difference in rating behavior between US and Non-US voters (FIG. 8, FIG. 9) [1].
- **Finding 4: Not All Metadata is Equally Predictive:** The paper's correlation analysis on "Parental Guide" items (e.g., "Violence & Gore") found an "almost-zero correlation" with ratings (FIG. 7) [1].
- **Finding 5: Temporal Genre Trends are Confirmed:** The paper explicitly notes a "descending trend of fantasy genre since 1994" and an "increasing the percentage of documentary... overtime" [1].

D. Identified Research Gaps

While foundational, the 2020 paper's primary value is in the research gaps it leaves open.

- **Gap 1: Simplistic Genre Analysis:** The paper's analysis is limited to basic statistics. It does not analyze the effect of *genre combinations* (e.g., Horror-SciFi-Comedy).
- **Gap 2: The "Individual" Analysis Failure:** The authors *failed* to meaningfully analyze the impact of "Director, Writers, Stars," stating the data is "some random string" and they "do not have a lot of data to assign them a value" [1].
- **Gap 3: Identified Bias, But No Solution:** The paper *proves* a measurable bias exists between US and Non-US voters but does not explore the *solution* (i.e., what makes a non-US film successful *despite* the bias).
- **Gap 4: Exclusive Focus on Movies:** The paper's analysis is entirely focused on movies, providing no insights into the quality patterns of TV Series or episodes.

III. Research Methodology

Our methodology is structured around six hypotheses, each corresponding to a research question. All methods are drawn from our data mining course syllabus.

A. Justification for Research Questions

Our 6 research questions are designed to be the innovative next steps that directly address the gaps identified in our literature review. The table below maps our questions to the justification.

TABLE I
Justification for Research Questions

Our Research Question	Justification based on Literature Review
Q1: Tracking Movie Genre Popularity	**Builds on Finding 5:** The paper confirms temporal genre trends exist. Our question formalizes this by creating a comprehensive visualization of both popularity and quality over time.
Q2: TV Show "Rating Decay"	**Fills Gap 4:** This is a new area of inquiry not covered by the paper. Our question extends the paper's temporal analysis methods to the unanalyzed TV Series dataset.
Q3: The "Genre Hybridity" Paradox	**Builds on Gap 1:** The paper's analysis stops at single-genre statistics. Our question explores the *interaction effects* of genre combinations.
Q4: The "Creative Hierarchy"	**Fills Gap 2:** The paper *gave up* on analyzing creative roles. Our question uses a superior methodology (Decision Trees) to rank the "Creative Hierarchy."
Q5: The "Cross-Cultural Breakout"	**Solves Gap 3:** The paper *identified* the problem (regional bias). Our question builds the *solution*: a predictive model to find the "fingerprints" of non-US films that overcome this bias.
Q6: The "High-Profile Flop"	**Fills Gap 2:** The paper's inability to quantify "high-value inputs" made this risk model impossible. By solving the "Star Power" problem, we can build this high-value model.

B. Foundational Hypotheses (H1–H2)

1) H1: Tracking Movie Genre Popularity:

- **Hypothesis (H1):** The popularity (median `numVotes`) and quality (median `averageRating`) of major movie genres show significant positive or negative trends over the last 30 years.
- **Null Hypothesis (H0):** There are no significant long-term trends; any changes are random or cyclical.
- **Methodology:**
  1) Filter the dataset for `titleType = ''movie''` and `startYear` (**Data Preprocessing**).
  2) Group the data by `startYear` and `primary_genre` (**Aggregation**).
  3) Calculate the `Median averageRating` and `Median numVotes` for each group (**Summary Statistics**).
  4) Plot these two metrics over time using a `Line Plot` (**Visualization**).

2) H2: TV Show "Rating Decay":

- **Hypothesis (H1):** Long-running TV series (e.g., 8+ seasons) show a significant negative trend (decay) in their `Mean averageRating` as the `seasonNumber` increases.
- **Null Hypothesis (H0):** There is no significant negative correlation between `seasonNumber` and `Mean averageRating`.
- **Methodology:**
  1) Identify a list of popular, long-running `tvSeries` from `title.basics`.
  2) Join with `title.episode` and `title.ratings` (**Data Preprocessing**).
  3) Group the data by series (`parentTconst`) and `seasonNumber` (**Aggregation**).
  4) Calculate the `Mean averageRating` for each

season (**Summary Statistics**).
5) Plot `seasonNumber` vs. `Mean averageRating` using a `Line Plot` (**Visualization**).

C. Innovative Hypotheses (H3–H6)

1) H3: The "Genre Hybridity" Paradox:

- **Hypothesis (H1):** We hypothesize that "Genre Purity" (a single genre tag) correlates with a higher `Mean averageRating` and lower `Variance` than "Genre Hybridity" (3+ genre tags).
- **Null Hypothesis (H0):** There is no significant difference in `Mean` or `Variance` between "Pure" and "Hybrid" films.
- **Methodology:**
  1) Engineer `genre_count` and `genre_type` features (**Feature Creation / Discretization**).
  2) Group the data by `primary_genre` (**Aggregation**).
  3) Calculate **Summary Statistics** (`Mean`, `Variance`) for both 'Pure' and 'Hybrid' films.
  4) Create side-by-side **Box Plots** (**Visualization**) to visually compare the groups.

2) H4: The "Creative Hierarchy":

- **Hypothesis (H1):** The `writer_tier` feature will be selected as the root split (highest **Gini Index** gain) for "plot-driven" genres (e.g., Mystery), while `director_tier` will be the root split for "spectacle-driven" genres (e.g., Action).
- **Null Hypothesis (H0):** The same feature will be the root split (best **Gini Index**) for all major genres.
- **Methodology:**
  1) Engineer `writer_tier`, `director_tier`, and `cast_tier` features (**Feature Creation**).
  2) Create a binary target `is_success` (e.g., `averageRating` > 7.0) (**Binarization**).
  3) Train separate **Decision Tree Classifiers** (**Decision Tree Induction**) for each major genre.
  4) Inspect the **root node** of each tree to identify the feature selected as the best split (based on the **Gini Index**).

3) H5: The "Cross-Cultural Breakout" Formula:

- **Hypothesis (H1):** A classifier can identify non-US "breakout" films with low **Classification Error**, and a **Decision Tree** will show that `genre` and `runtimeMinutes` are the most predictive features.
- **Null Hypothesis (H0):** A classifier will perform no better than a random baseline, and the **Classification Error** will be high.
- **Methodology:**
  1) Create the binary target `is_breakout` (**Feature Creation / Binarization**).
  2) Train a **Random Forest Classifier** (**Ensemble Methods**) to predict this target.
  3) Address the **Class Imbalance Problem** using techniques like **Sampling**.
  4) Evaluate the model using **k-fold Cross-Validation** and its average **Classification Error**.

4) H6: Predicting the "High-Profile Flop":

- **Hypothesis (H1):** A classifier can identify "High-Profile Flops" with a low **Classification Error** for the minority 'flop' class. We hypothesize that `genre` and `runtimeMinutes` will be key predictors.
- **Null Hypothesis (H0):** No metadata features can reliably predict a flop given high-value inputs; the **Classification Error** will be high.
- **Methodology:**
  1) Engineer binary features `is_high_profile` and `is_flop` (**Feature Creation / Binarization**).
  2) Train and compare classifiers such as a **Naive Bayes Classifier** and a **Random Forest Classifier**.
  3) Use techniques to address the severe **Class Imbalance Problem**.
  4) Evaluate the models using **Cross-Validation** and **Classification Error**.

IV. Data Pre-processing Plan

A robust, syllabus-compliant pipeline is required to execute our methodology.

A. Source Datasets

We will use the datasets as described in Table II.

TABLE II
Source Datasets and Key Fields Used

Dataset	Description	Key Fields Used
`title.basics`	Core metadata	`tconst`, `titleType`, `startYear`, `runtimeMinutes`, `genres`
`title.ratings`	User ratings	`tconst`, `averageRating`, `numVotes`
`title.episode`	TV episode links	`tconst`, `parentTconst`, `seasonNumber`
`title.crew`	Writers/Directors	`tconst`, `directors`, `writers`
`title.principals`	Cast/Crew	`tconst`, `nconst`, `category`
`name.basics`	Person info	`nconst`, `primaryName`
`title.akas`	Regional info	`titleId`, `region`, `language`, `isOriginalTitle`

B. Data Cleaning and Standardization

We will create two separate data flows:

- **Flow 1: Movie Dataset (Q1, 3, 4, 5, 6):** Filter `title.basics` for `titleType = ''movie''`, `startYear` ≥ 1990, and `numVotes` ≥ 100.
- **Flow 2: TV Episode Dataset (Q2):** Filter `title.basics` for `titleType =`

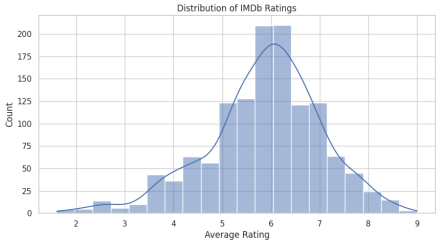

Fig. 1. Distribution of IMDb Ratings.

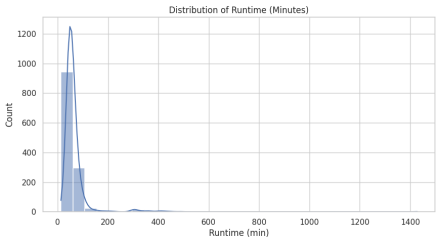

Fig. 2. Distribution of Runtime (Minutes).

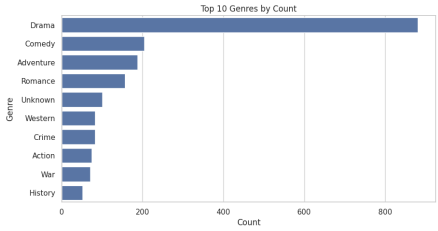

Fig. 3. Top 10 Genres by Movie Count.

`''tvSeries''`, then join with `title.episode` and `title.ratings`.

For handling **Missing Values**, `runtimeMinutes` will be imputed with the `Median` value of its genre (**Summary Statistics**). Initial exploratory plots, such as those in Fig. 1-3, confirm the distributions of key features.

C. Data Integration Steps (*Aggregation*)

To prevent a "row explosion," all one-to-many joins will be handled via aggregation.

- **Movie Pipeline:** We will pre-process `title.akas` to get one primary `region` per movie. We will process `title.principals` and `title.crew` separately to create aggregated features (e.g., `director_tier`) *before* joining them to the main `movies_df`.
- **TV Pipeline:** We will join `title.episode` with `title.ratings` on the episode's `tconst`.

D. Feature Engineering

The features in Table III will be engineered to support our hypotheses.

TABLE III
FEATURE ENGINEERING PLAN

Feature	Syllabus Technique	Relevant Question(s)
`primary_genre`	**Feature Creation**	Q1, Q3
`genre_count`	**Feature Creation**	Q3
`genre_type`	**Discretization**	Q3
`director_tier`	**Discretization**	Q4, Q6
`writer_tier`	**Discretization**	Q4
`cast_tier`	**Discretization**	Q4, Q6
`is_success`	**Binarization**	Q4
`is_breakout`	**Binarization**	Q5
`is_flop`	**Binarization**	Q6
`normalized_votes`	**Variable Transformation**	Q5

E. Final Datasets

Our pipeline will generate modular datasets, one for each hypothesis, as shown in Table IV.

TABLE IV
DATASET SPLITTING FOR HYPOTHESIS TESTING

Dataset	Target Hypothesis
`df_genre_trends.csv`	Q1: Movie Genre Trends
`df_tv_decay.csv`	Q2: TV Show Rating Decay
`df_genre_hybridity.csv`	Q3: Genre Hybridity
`df_creative_hierarchy.csv`	Q4: Creative Hierarchy
`df_cross_cultural.csv`	Q5: Cross-Cultural Breakout
`df_flop_model.csv`	Q6: High-Profile Flop

V. CONCLUSION AND FUTURE WORK

A. Conclusion

The Bahraminasr (2020) paper is the ideal academic foundation for our project. It validates our basic premises (e.g., metadata is predictive, regional bias exists) while leaving clear, high-value research gaps that our project is designed to fill. The paper's focus on a demographic dataset (age/gender) which we lack, combined with its failure to analyze individuals (actors/directors), gives our project a clear and innovative focus. We will leverage the official IMDb datasets to conduct the sophisticated, role-based metadata analysis that this paper proved was a missing piece of the puzzle.

B. Future Work

This paper has presented a complete and rigorous plan for Phase 1 of our project. The immediate next steps are to execute this data pre-processing pipeline to generate the modular datasets described in Table IV. Following this, we will proceed to Phase 2, conducting the Exploratory Data Analysis (EDA) and visualizations as defined in our `eda_and_visualization_plan.md`. Finally, we will enter Phase 3, training and evaluating the classification models (H4, H5, H6) as outlined in our `model_training_plan.md` to test our predictive hypotheses.

REFERENCES

[1] A. 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