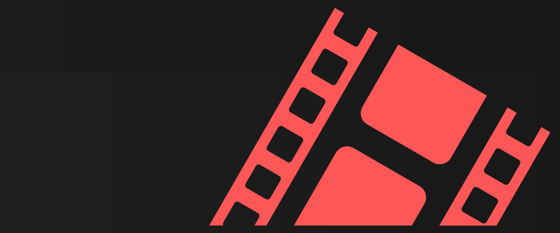
From Data to Blockbusters: Predicting the Next AAA Movie Hit



AAA MOVIE PREDICTION

AUSTINE WONG

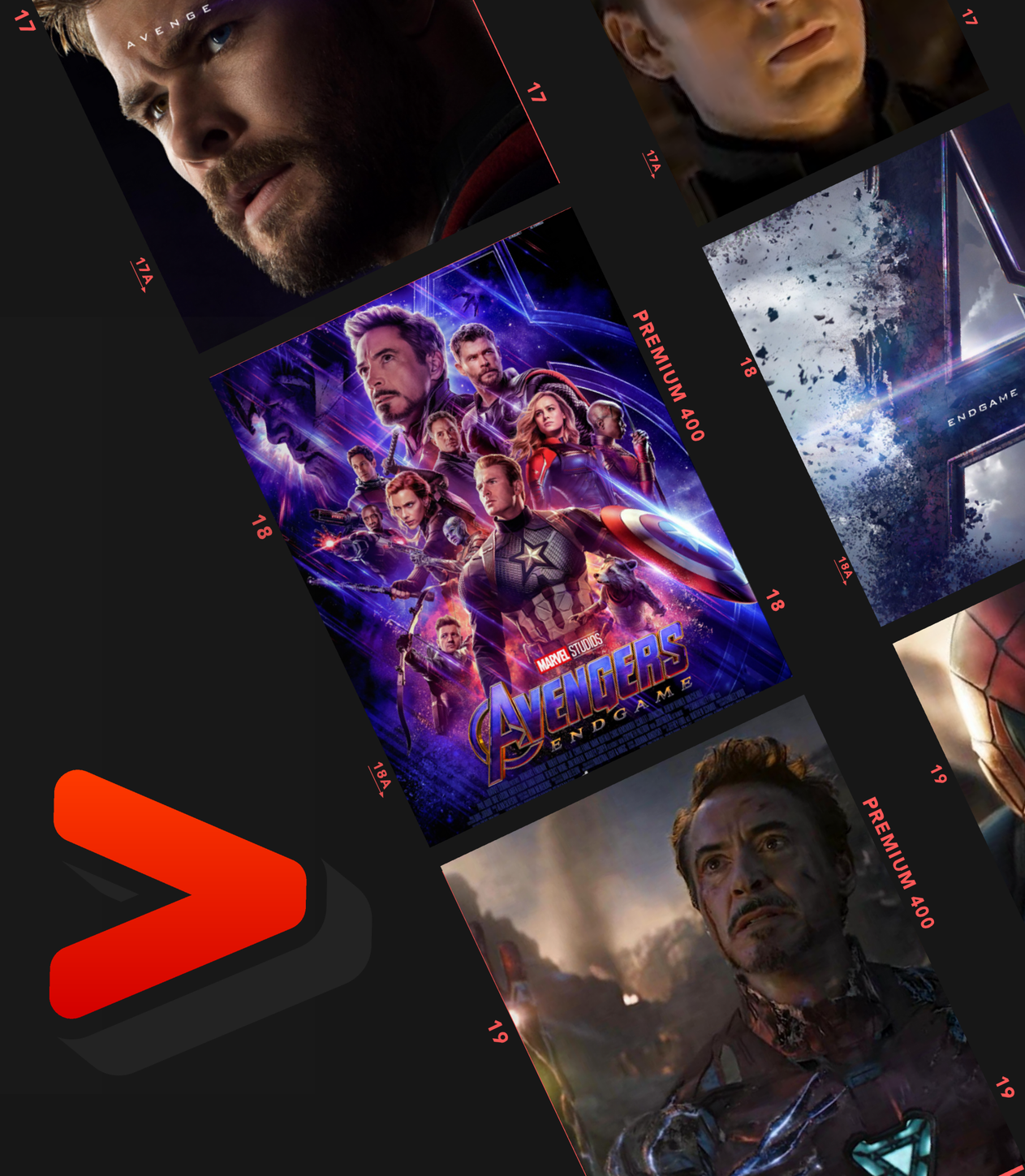
PREDICT THE NEXT AAA MOVIE

What makes a movie memorable?

Some movies stand out, captivating audiences and critics alike. Think of Avengers: Endgame—a blockbuster that not only smashed box office records but also left an enduring cultural impact.

Yet, not every movie achieves this level of success. Studios spend millions crafting films, often without a clear understanding of their potential reception.

The Challenge:
How can we predict and shape the next AAA title?
Is it possible to use data to remove guesswork and maximize the chances of producing a hit?

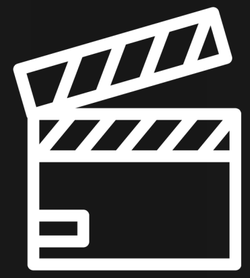


Our Goal

Feature Importance

Identify trends and key success factors.

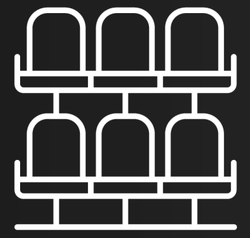
EDA



Predict

Leverage the IMDb Datasets to shape and predict the next AAA title.

ML



Explain

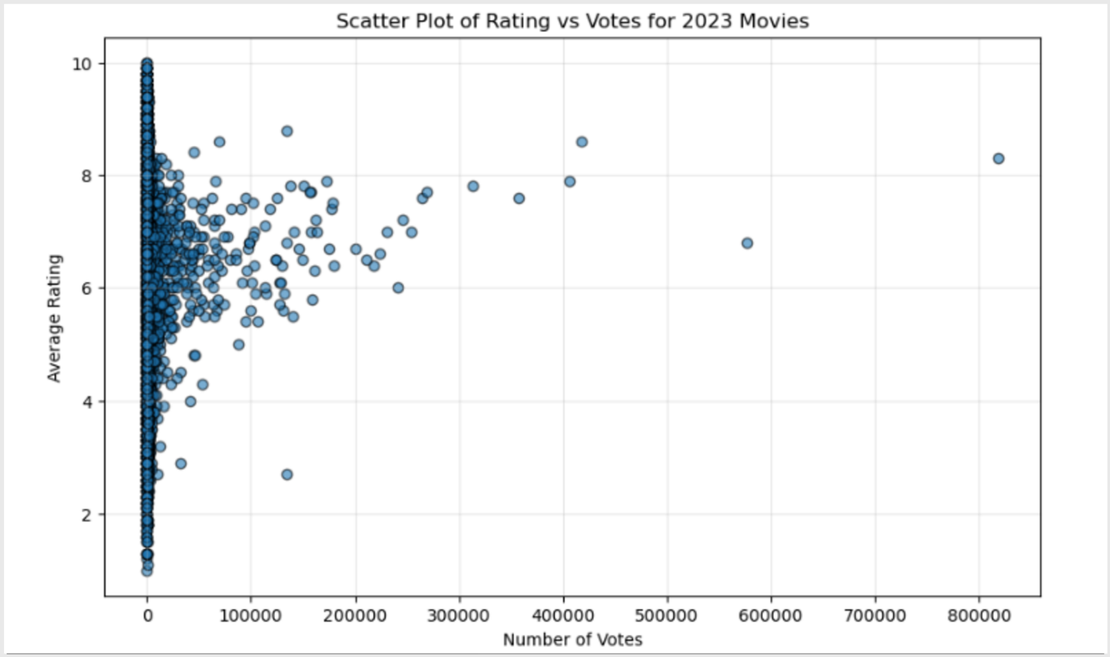
Help movie studios make informed, data-driven decision

SHAP

IMDb Non-Commercial Datasets

IMDB provides access to publicly available data about movies, TV shows, actors, and other entertainment-related entities. The datasets are structured to allow easy analysis and integration into machine learning or data science projects.

Dataset File	Description	Key Columns
title.basics.tsv	Contains basic information about titles.	tconst (unique identifier), titleType, primaryTitle, originalTitle, isAdult, startYear, genres
title.crew.tsv	Details about directors and writers associated with titles.	tconst, directors (comma-separated), writers (comma-separated)
title.principals.tsv	Information about the main cast and crew of a title.	tconst, ordering, nconst (individual identifier), category, job, characters
title.ratings.tsv	User ratings and aggregated scores for titles.	tconst, averageRating, numVotes
name.basics.tsv	Biographical information about individuals.	nconst (unique identifier), primaryName, birthYear, deathYear, primaryProfession, knownForTitles



What was 2023's fan favorite?

Movies with higher ratings tend to cluster with fewer votes, suggesting niche appeal.

A select few movies balance high ratings with a significant number of votes, indicating mass appeal (AAA movie traits).

How do we balance ratings and number of votes?
Candidates for AAA status often have consistent quality and wide-reaching impact.

PREDICT THE NEXT AAA MOVIE

What was 2023's fan favorite?

Movie	Rating	NumVotes	Weighted Score	Genre
Spider-Man: Across the Spider-Verse	8.6	417620	8.45	Action, Adventure, Animation
12th Fail	8.8	134765	8.36	Biography, Drama
Oppenheimer	8.3	817891	8.23	Biography, Drama, History

This approach balances ratings and number of votes, ensuring movies with broader audience feedback rank higher, making the results more reliable and representative.

Weighted Score Formula

Derived from a common technique used to **balance popularity and quality** in rankings. It's inspired by methods like the Bayesian average and IMDb's own rating calculation, designed to adjust a movie's rating based on the number of votes it has received

$$WS = \frac{(\#\text{of Votes} * \text{Rating}) + (\text{minimum Votes} * \text{Global Average})}{\#\text{ of Votes} + \text{minumum Votes}}$$

minimum Votes = 25,000

- Consistency with IMDb)
- Filters out movies with insufficient votes, ensuring the list reflects movies that are both highly rated and widely voted upon.

- **Fairness**: Movies with fewer votes are not overly penalized but also do not dominate the rankings.
- **Reliability**: Ratings with many votes are weighted more heavily, making them statistically more trustworthy.

Who captured the spotlight?

Issa Rae

1.09 M

Barbie

★ 6.8
576 K

Jessica Drew

★ 8.6
418 K

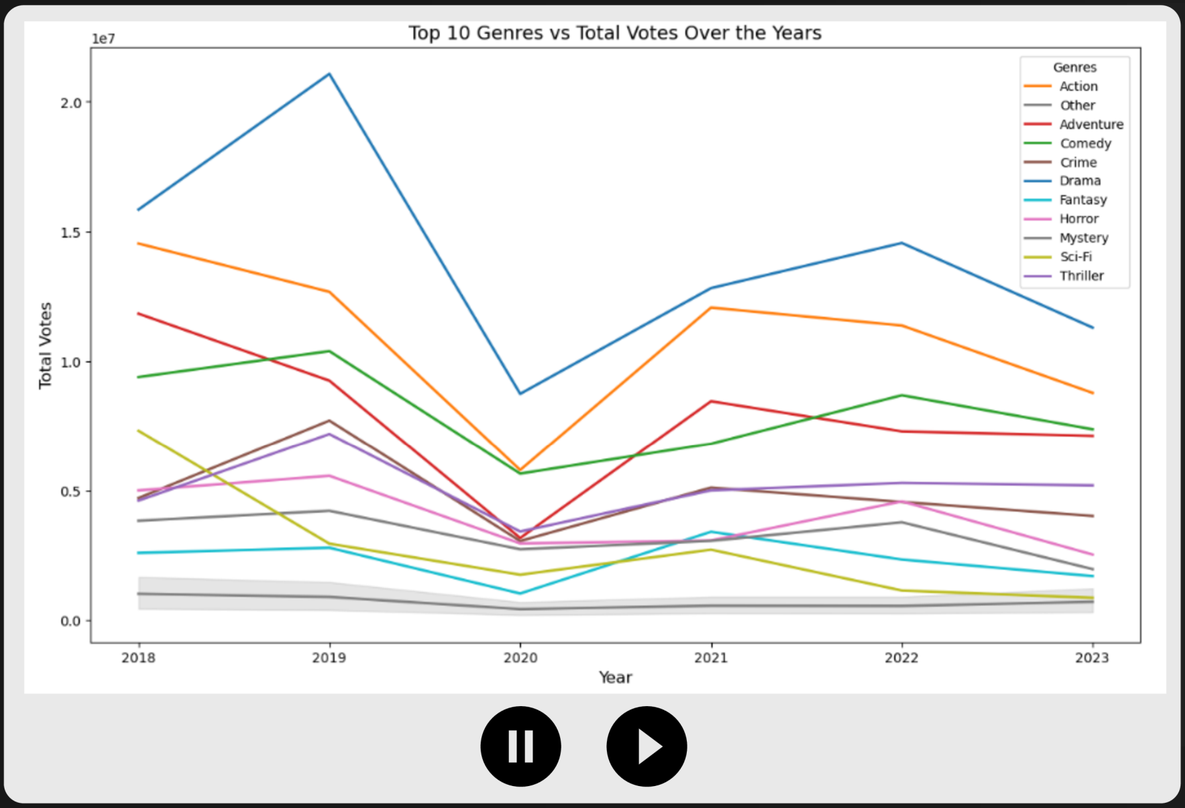
Sinatra Golden

★ 7.5
102 K

Most Popular Actor in 2023 by Total Votes (User Interaction)

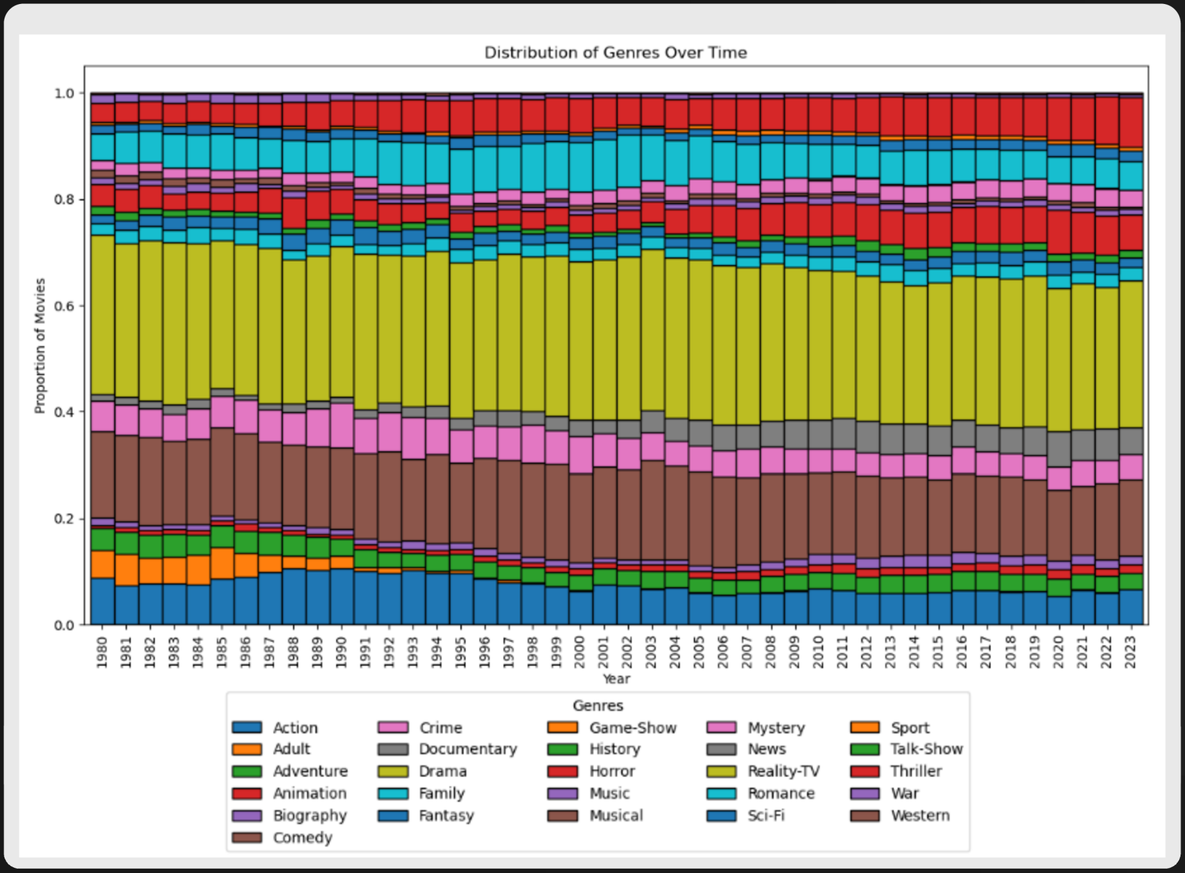
In 2023, **Issa Rae** stood out as the most popular actor based on **total votes**, **high-quality movie appearances**, and her involvement in projects spanning **diverse genres**. Her ability to be part of widely successful movies like Spider-Man: Across the Spider-Verse further solidified her status as a top actor for the year.

Issa Rae (1.09M; 3 Movies) | Chris Pratt (1.06M, 3 Movies) | Matt Damon (995K; 2 Movies)



Trends in User Preferences by Genre (2018–2023)

- **Action** and **Adventure** continue to **dominate**, proving their resilience and reliability for studios aiming for large-scale appeal.
- **Sci-Fi's stagnation** after 2021 indicates a need for fresh narratives or cross-genre blending to reignite audience interest.
- **Niche genres like Fantasy and Mystery** could be emerging opportunities, appealing to specific audience demographics post-pandemic.



Genre Distribution Trends Over Time

- The **increasing share of Action and Adventure** highlights a shift toward high-budget, spectacle-driven filmmaking.
- The **continued dominance of Drama** indicates its versatility and universal appeal, though it may require hybridization with popular genres (e.g., action-drama).
- **Comedy** also remains a major genre, although its share has gradually declined in recent years.

PREDICT THE NEXT AAA MOVIE

Defining the Metric for a Hit Movie

AAA Movie Score (AMS)

Normalized Box Office Revenue

Studios prioritize profitability, and this ensures fair comparisons across all projects.

CHOSEN

Optimized Rating

Ensures investment in movies with quality-driven appeal.

Social Media Engagement Score

High engagement on platforms indicates strong audience interest, critical for modern marketing strategies.

The AMS metric is a decision-support tool for studio directors, blending profitability, quality, and buzz into a single actionable framework. By focusing on measurable predictors, it minimizes uncertainty and maximizes the likelihood of producing the next blockbuster.

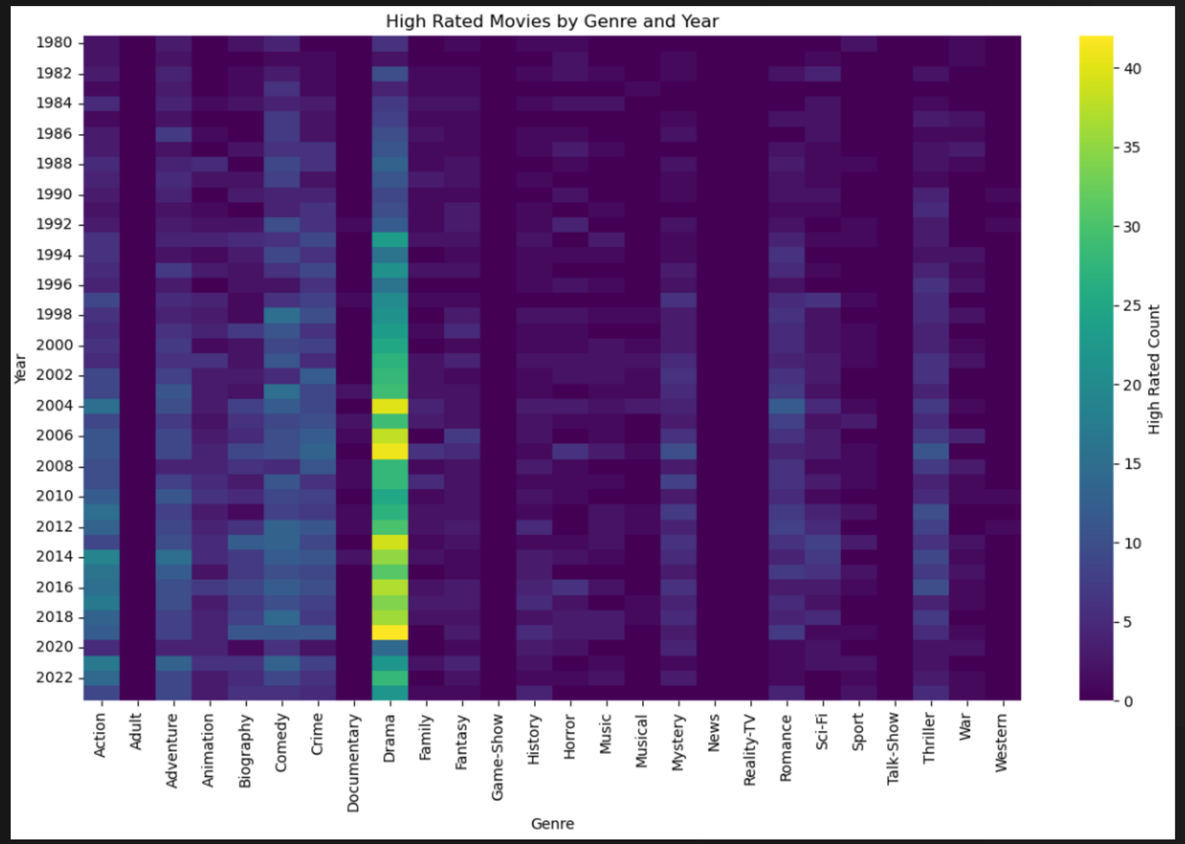
PREDICT THE NEXT AAA MOVIE

AAA Title Rating Threshold

A small, highly selective group that consistently delivers both critical and audience acclaim. This approach ensures our analysis focuses on films with proven success metrics, aligning with the goals of predicting the next triple-A title.

0.8%

Minority
Selecting movies with weighted average of 7.0 or higher



High Rated Movies by Genre and Year

- **Drama** has the highest concentration of high-rated movies over time, particularly between 2000 and 2020.
- **Documentary** consistently contributes a significant number of high-rated movies, especially from 2005 onward.
- **Biography** shows a steady increase in high-rated movies over recent years.
- Certain genres, like **Crime**, **Thriller**, and **Horror**, have sporadic contributions, suggesting they may be less dominant overall but could be important for niche predictions.

Data Preprocessing and Feature Engineering

Original Features
tconst
primaryTitle
startYear
runtimeMinutes
genres
averageRating
numVotes
actors
directors
writers

Feature Engineer	Description
weightedScore	balancing averageRating and numVotes
HighRated	weightedScore >= 7
prior_weightedScore_actor	prior average weightedScore
prior_weightedScore_director	prior average weightedScore
prior_weightedScore_writer	prior average weightedScore
prior_highRated_actor	prior count of highRated movies
prior_highRated_director	prior count of highRated movies
prior_highRated_writer	prior count of highRated movies

One-Hot Encode
genres (26)

Features Used for Predictive Modelling

Target Variable

Motivations

	Auto Machine Learning Pipelines	XGBoost with SHAP	GNNs
Purpose	**AutoML Frameworks (AutoKeras & Auto-sklearn)** • Reduces human intervention by systematically searching over model and parameter spaces, ensuring efficiency and reproducibility.	**XGBoost with SHAP** • Provides a highly efficient predictive model while maintaining transparency and interpretability.	**Graph Neural Networks** • Learns hierarchical representations of graph-structured data, capturing dependencies between entities (e.g., actors, genres, directors in movies).
Concerns	• Faced installation issues and dependency conflicts. • Difficult to customize for my specific dataset.		• High computational complexity and steep learning curve. • Limited practicality within time and resource constraints.
Solution	**Custom AutoML Pipeline** • Built a lightweight pipeline with **Random Forest** and **XGBoost**. • **XGBoost** selected as the final model for its superior performance. • Integrated **SHAP** for feature interpretability, ensuring a balance of performance and explainability.		

Predictive Modelling

Train Validation Holdout Split

Dataset Split	non-HighRated	HighRated	Total Count
Train Validation	130,152	1,049	131,201
Holdout	43,384	350	43,734

Auto-ML

def auto_ml(X, y, models_dict, scaler=None, cv=None, res_t=None):

Models
- Random Forest
- XGBoost

Scaler
- MinMaxScaler

CV
- StratifiedKFold

Resampling
- SMOTE
- ADASYN

Predictive Modelling

Base Model

Model	Train Accuracy	Val Accuracy	Train Recall	Val Recall	Train Precision	Val Precision	Train f1	Val f1	Run Time
Random Forest	100.00%	99.33%	100.00%	29.74%	100.0	69.3	100.0	41.58	8.736971
XGBoost	99.90%	99.32%	88.49%	42.42%	98.75	60.33	93.34	49.8	1.095757

Why focus on Recall?

- Minimizing Missed Opportunities
- Strategic Resource Allocation
- Creative Enhancements

High recall prioritizes identifying nearly all AAA movie candidates, aligning with the industry's goal of maximizing opportunities and minimizing the risk of missing a potential blockbuster.

Predictive Modelling

Random Forest

	Method	Train Recall	Val Recall	Holdout Recall
0	No Resampling	100.00%	29.74%	31.71%
1	Resampling	100.00%	76.55%	77.71%
2	Resampling x Hypertune	99.67%	96.95%	97.14%

XGBoost

	Method	Train Recall	Val Recall	Holdout Recall
0	No Resampling	88.49%	42.42%	47.14%
1	Resampling	97.62%	74.26%	77.43%
2	Resampling x Hypertune	98.29%	98.67%	99.71%

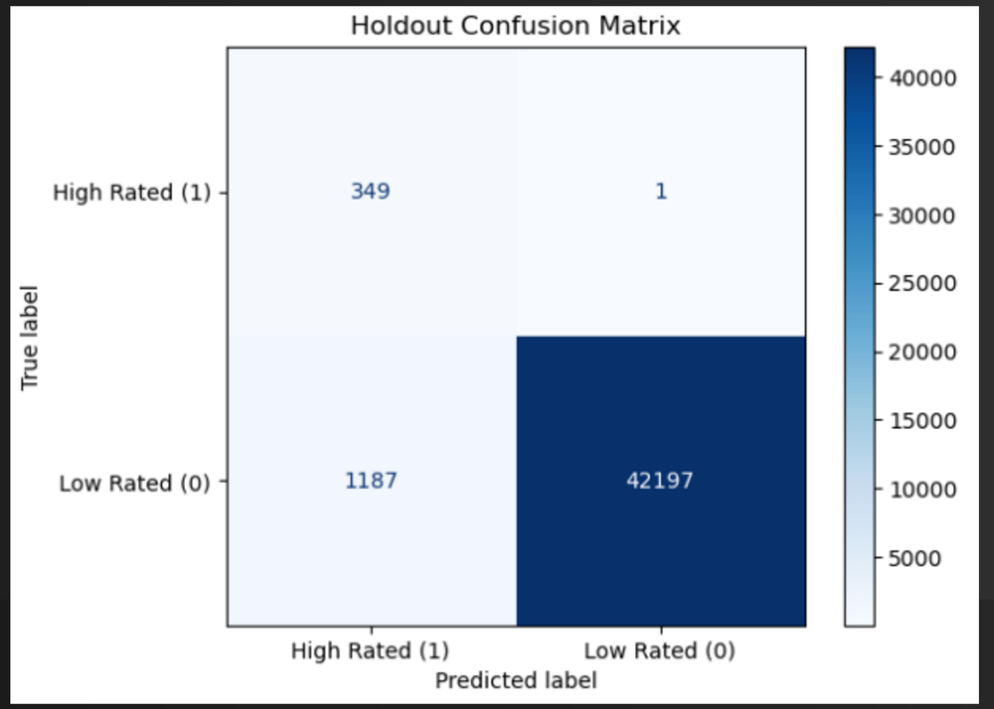
Predictive Modelling

XGBoost

	Method	Train Recall	Val Recall	Holdout Recall
0	No Resampling	88.49%	42.42%	47.14%
1	Resampling	97.62%	74.26%	77.43%
2	Resampling x Hypertune	98.29%	98.67%	99.71%

- The model successfully captures almost all AAA titles, with a recall of ~99.7% (only 1 out of 350 actual AAA movies is missed). This means the model is very reliable at ensuring AAA titles are identified.
- The model predicts many non-AAA (low-rated) movies as AAA (1,187 false positives), leading to a precision of ~22.7% for high-rated movies. This implies that while most AAA titles are captured, the predictions include a significant number of false positives.

Confusion Matrix (Unseen Data)



Combining XGBoost with SHAP (SHapley Additive exPlanations)

XGBoost and **SHAP** together create a powerful framework for predictive modeling and interpretation. XGBoost, a gradient boosting algorithm, **excels at making accurate predictions**, while SHAP provides **i**nterpretability by **explaining the contributions of each feature** to the predictions. This combination is especially impactful in domains like predicting AAA movie titles, where complex relationships and high-stakes decision-making require both accuracy and transparency.

[1] Jabeur, S.B., Mefteh-Wali, S. & Viviani, JL. Forecasting gold price with the XGBoost algorithm and SHAP interaction values. Ann Oper Res 334, 679–699 (2024). https://doi.org/10.1007/s10479-021-04187-w

[2] Schoonemann, J., Nagelkerke, J., Seuntjens, T.G. et al. Applying XGBoost and SHAP to Open Source Data to Identify Key Drivers and Predict Likelihood of Wolf Pair Presence. Environmental Management 73, 1072–1087 (2024). https://doi.org/10.1007/s00267-024-01941-1

[3] Yi, F., Yang, H., Chen, D. et al. XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease. BMC Med Inform Decis Mak 23, 137 (2023). https://doi.org/10.1186/s12911-023-02238-9

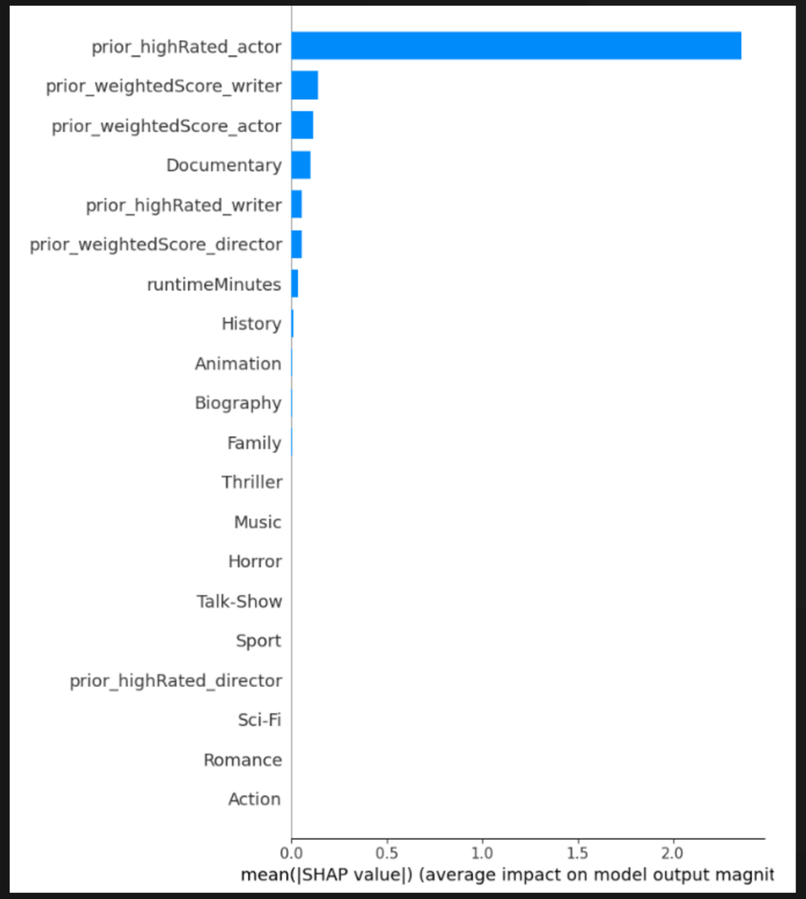
Recent Applications:

Financial Forecasting:
In 2021, researchers utilized XGBoost alongside SHAP to enhance gold price forecasting. The study demonstrated that this combination **not only improved predictive performance but also provided clear insights into the influence of different features** on gold price movements. [1]

Environmental Management: A 2024 study applied XGBoost with SHAP to predict wolf pair presence in Germany. The model achieved a high Area Under the Curve (AUC) of 0.91, with SHAP analysis **identifying key drivers such as proximity to neighboring wolf pairs and percentage of wooded area.** [2]

Healthcare Diagnostics: In 2023, an interpretable diagnostic framework for Alzheimer's disease was developed using XGBoost and SHAP. This approach allowed for **accurate predictions while elucidating the contribution of various biomarkers** to the diagnosis. [3]

Combining XGBoost with SHAP (SHapley Additive exPlanations)



Feature Importance:

Top Features:

- High-Rated Actor:
  - Strongest predictor
  - actors with a history of high-rated movies heavily influence AAA predictions.
- Writer Success
  - Proven writers impact predictions but less than actors.
- Documentary Genre
  - Significantly affects predictions due to niche scoring trends.

Moderate Influence:

- Success of directors, writers, and other actors moderately affects predictions.

Low Influence:

- Features like runtime and genres (History, Biography, Romance, etc.) have minimal impact.

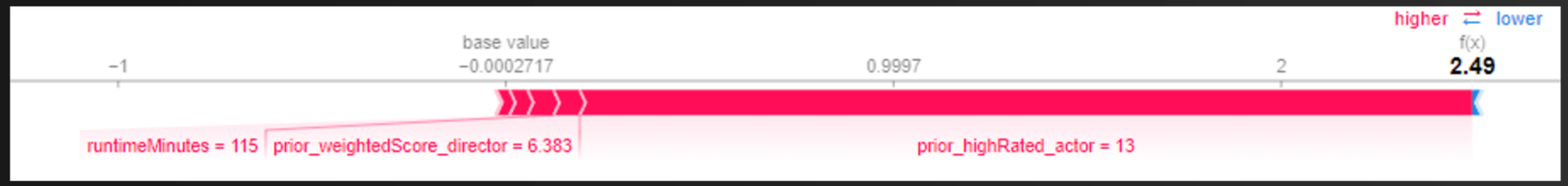
Testing on Unseen Data

tick, tick... BOOM!

⭐ **7.5**
122 K

XGBoost Prediction:
AAA ✓

SHAP FORCE PLOT



This chart shows how the feature prior_highRated_actor affects the prediction for whether this movie could be the next AAA title.

The base value is about **-0.00027**, which is the average prediction the model makes before looking at any features. The final prediction for this movie is 2.49, and the main reason for this increase is the feature prior_highRated_actor, which is 13 here.

What does that mean?

- It means the actors in this movie have been in 13 highly rated films before.
- The model sees this as a very positive sign, because actors with strong track records often bring credibility and fan interest, which are key for predicting a movie's success before release.
- So, in simple terms, the model predicts a high chance of success partly because of the strong cast involved in the movie.

Business Value

- Minimizing Missed Opportunities
- Building Trust
- Strategic Resouce Allocation
- Informed Decision Making

Using **XGBoost's high recall capabilities** and **SHAP's interpretability**, we can **minimize missed opportunities** in identifying AAA movies, ensuring no high-potential titles are overlooked.

This approach maximizes business value by capturing all promising candidates for targeted marketing, investment, and **strategic resource allocation**.

While XGBoost ensures accurate predictions by capturing complex patterns, SHAP explains the key drivers behind these predictions—such as the influence of high-rated actors or directors—**building trust** and enabling informed decision-making.

Together, they reduce the risk of missing blockbusters, optimize ROI, and align production strategies with actionable insights.

Future Improvements and Recommendations

While the current XGBoost + SHAP framework effectively predicts AAA titles, exploring Graph Neural Networks (GNNs) offers a promising avenue for enhancing predictive power. Relational data, central to movie success, can be better modeled through GNNs.

Why GNNs?
GNNs excel in capturing relationships and patterns within graph-structured data, making them ideal for movie analytics.

Integrating GNNs with the existing model ensures a more holistic understanding of relational data, making predictions more accurate. As studios rely on increasingly complex datasets, adopting GNNs will offer a significant competitive edge in identifying the next AAA blockbuster.

PREDICT THE NEXT AAA MOVIE

THANK YOU

Austine Rico Wong