Movies: What Drives Audience Enjoyment

1. Problem Justification and Relevance

Choosing a movie may seem simple, yet audiences differ greatly in how they evaluate films. Some rely on critic platforms like IMDb or Rotten Tomatoes, while others focus on attributes such as genre, cast, or budget. Understanding this gap matters because it determines how much trust viewers can place in critical scores and how effectively these platforms guide audiences toward enjoyable choices. We selected this topic because choosing a movie is a personal and subjective experience, and many viewers struggle to decide whether to simply trust critics or if there is a better method of evaluating movie quality. By analyzing both perspectives, we aim to provide clarity on what truly shapes an audience's enjoyment. Our goal is to help viewers make more informed decisions based on evidence rather than assumptions. Therefore, our analysis is centered around the following questions:

- Which movie critic platform provides a more reliable guide for audiences?
- Does the more reliable movie critic platform do an accurate job of predicting audience enjoyment?
- How can understanding factors outside of general critic ratings help future audiences make even more informed movie watching decisions?

2. Related Work

A key paper that influenced our approach was *Predicting Movie Ratings Based on Metadata* (Chen et al., 2024). The authors used pre-release metadata, such as director quality, actor score, genre, and budget, to predict IMDb ratings. Their Random Forest model performed strongly (R^2 up to 0.909, RMSE ≈ 0.48), and they found that director quality was the most influential factor. Their emphasis on feature engineering and structured metadata informed us of our decision to use a similar multi-step modeling approach. Another study that shaped our project was *A Machine Learning Approach to Predict Movie Revenue Based on Pre-Released Metadata* (Mahmud et al., 2020). The paper explored models such as Logistic Regression, SVM, MLP, and Decision Trees to predict film profitability. Their work highlighted the high uncertainty of movie success where only 36% of films between 2000 and 2010

recovered their production budgets. This reinforced the motivation for our project, which also relies on metadata and attempts to model outcomes before audience reactions are fully known.

3. Data Acquisition and Description

3.1 Data Source & Collection

We collected two datasets from Kaggle, IMDB and Rotten Tomato. These datasets were downloaded in their original CSV format and then merged to create a unified dataset. It contains our target variable, audience ratings, and features that can be used to predict audience ratings such as critical text sentiment and movie budget.

3.2 Data Preprocessing

SQL:

Preprocessing Step	Justification
Clean text for director and movie title key columns	Allow for smooth merging of datasets and to allow for future usage of the variables for analysis
Merge IMDb and Rotten Tomatoes datasets	Create a single dataset for analysis that combines unique features available in both datasets
Select necessary columns for analysis (ex: budget and run time)	Remove redundant features in the dataset from the table merging

Python:

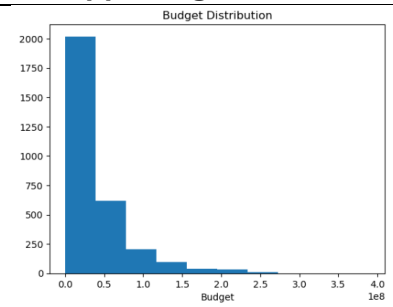
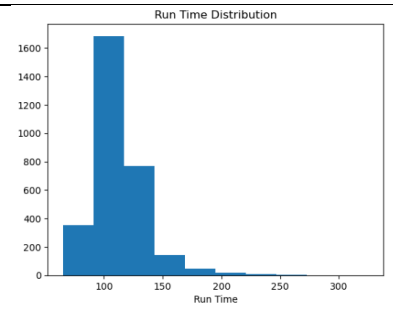
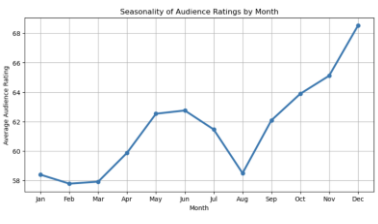
Preprocessing Step	Justification
Convert necessary columns from strings to numeric data	Make all features usable for machine learning methods
Fill three missing audience rating values with median audience rating value	Allows for all datapoints in the dataset to be usable in analysis

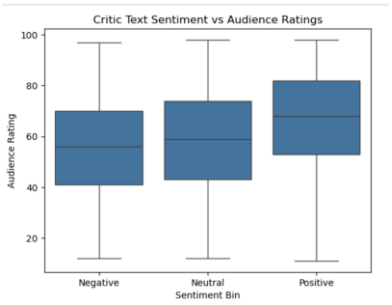
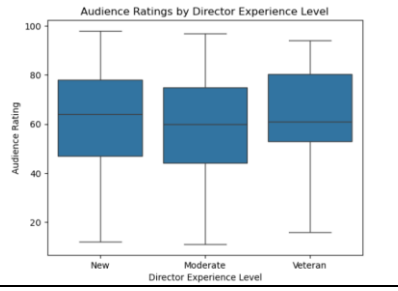
4. Data Exploration

4.1 Summary Statistics (Pre-Data Engineering)

	Running time	Budget	Actors Box Office %	Director Box Office %	IMDb score	tomatometer_rating	audience_rating
count	3021.000000	3.021000e+03	3.021000e+03	3021.000000	3021.000000	3021.000000	3021.000000
mean	111.476663	3.768899e+07	3.926220e+09	53.225746	6.489308	54.658060	61.790467
std	22.076025	4.160991e+07	1.542206e+11	33.737305	1.040899	27.577319	19.467993
min	65.000000	1.500000e+04	0.000000e+00	0.000000	1.600000	0.000000	11.000000
25%	97.000000	1.000000e+07	3.333000e+01	33.330000	5.900000	31.000000	47.000000
50%	107.000000	2.500000e+07	5.714000e+01	50.000000	6.600000	56.000000	63.000000
75%	121.000000	5.000000e+07	8.333000e+01	80.000000	7.200000	80.000000	78.000000
max	325.000000	3.900000e+08	6.805556e+12	100.000000	9.300000	100.000000	98.000000

4.2 Data Engineering

Feature Engineered	Reasoning	Supporting Visualization
Log Transformed Budget	The budget variable was highly right skewed, so to normalize the variable a log transformation was used.	
Log Transformed Run Time	The run time of movies was right skewed, to normalize this a log transformation was used on the column.	
Seasonality Feature	We noticed a relationship between the timing of a movie release and the audience reception of the movie.	

Text Sentiment Score	Our dataset contained text sentiment from critics; to make use of this column VADER sentiment scoring was used, and the scores were converted to sentiment bins (Negative, Neutral, Positive) to capture non-linear relationships.	
Historical Director Data	To capture the impact of director experience, a column for the number of films historically released in our dataset was engineered and converted to bins to capture non-linear relationships.	

5. Data Analysis and Results

5.1 Model Descriptions

Model	Description	Justification
Linear Regression	Fits a linear relationship between variables, minimizing squared errors.	Used as a baseline model
Polynomial Regression	Introduces polynomial terms to capture non-linear relationships between variables.	Used to model non-linear trends while retaining interpretability.
Polynomial Ridge Regression	Applies regularization to polynomial regression to reduce overfitting from complex features.	Used to stabilize polynomial models and control coefficient magnitude.
Decision Tree Regression	Use feature-based decision rules based on information gain to segment data into predictive regions.	Used for interpretability and capturing non-linear interactions.
Random Forest Regression	Aggregates multiple decision trees built on random samples to improve generalization.	Used to reduce overfitting and improve predictive performance compared to individual decision trees.
KNN Regression	Predicts outcomes by taking the average value of the target variable of nearby observations,	Used to capture local patterns without assuming a functional form.

	calculated by Euclidian distance in the feature space.	
XG Boost Regression	Sequentially builds boosted decision trees that correct prior errors using gradient optimization.	Used to capture complex non-linear relationships and maximize predictive performance.

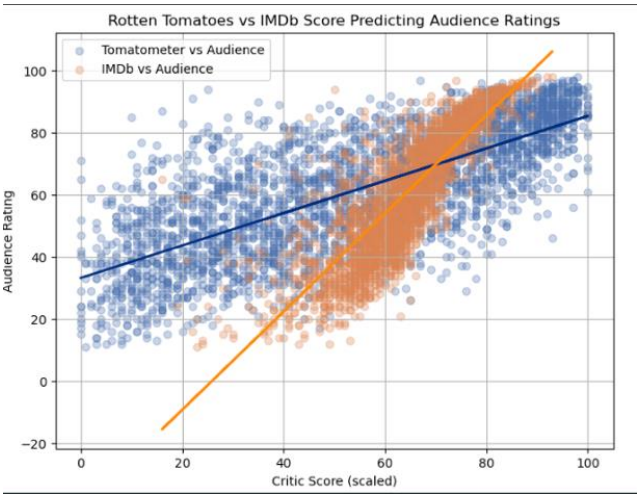
5.2 Analysis Part 1: IMDb vs. Rotten Tomatoes

Question to Answer: Which movie critic platform provides a more reliable guide for audiences?

Analytical Approach: Applied 5-fold cross-validated linear regression, using IMDb ratings and Rotten Tomatoes scores as predictors of audience ratings.

Key Results:

Rating Platform	Average R^2	MSE
Rotten Tomatoes	.544	10.51
IMDb	.714	7.75



Interpretation: IMDb ratings show stronger predictive reliability, explaining more variance and achieving lower error than Rotten Tomatoes. The tighter linear relationship and reduced dispersion indicate more consistent alignment with audience ratings.

Next Steps: Following the identification of IMDb as the stronger predictor of audience sentiment, we focus our next analysis exclusively on IMDb scores and optimize multiple machine learning models to enhance prediction accuracy.

5.3 Analysis Part 2: Optimized IMDb Model

Question to Answer: Does IMDb accurately predict audience enjoyment?

Analytical Approach: Apply multiple optimized models using IMDb as the sole predictor to assess its predictive power for audience ratings.

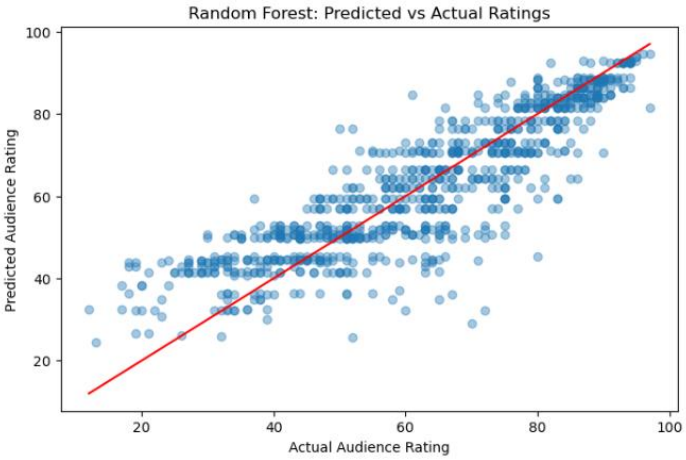
Key Results:

Model	Method	Best Parameters	Best R^2
Linear Regression	5-Fold Cross Validation	N/A	0.714
Polynomial Regression	Grid Search *degrees: 2, 3, 4*	Degree = 4	0.768
Decision Tree Regression	Grid Search *max depth: 1 - 6*	Max Depth = 5	0.770
Random Forest Regression	Grid Search *max depth: None,1,3,5,7* *estimators: 100 - 1100*	Max Depth = 5 Num Estimators = 600	0.771
KNN Regression	Grid Search *neighbors: 1 - 9*	N Neighbors = 8	0.747
XG Boost Regression	Grid Search *max depth: 1 - 11* *learning rate: 0.01 - 0.11* *estimators: 200 - 800*	Max Depth = 1 Learning Rate = 0.06 Num Estimators = 600	0.770

Model Selection: Random Forest Regression achieved the highest average R^2 (0.771) and was selected as the final model.

Model Assumptions: The selected Random Forest model uses a maximum depth of 5 and 600 estimator which limits model complexity and improves generalization. The shallowness of the trees prevents overfitting while the large number of trees generated results in increased stability as the results are based on an ensemble of 600 trees.

Interpretation: IMDb ratings alone demonstrate strong predictive capability, explaining 77.1% of the variance in audience scores using an optimized Random Forest model. The close alignment between predicted and actual audience ratings in the visualization confirms a strong positive relationship, indicating that IMDb provides an accurate signal of audience sentiment.



Next Steps: To further improve predictive performance and support informed movie-watching decisions, we incorporated engineered features, dummy variables, and additional model refinements to develop a more robust predictive model for audience ratings.

5.4 Analysis Part 3: Optimized Predictive Model

Question to Answer: How can understanding factors outside of general critic ratings help future audiences make even more informed movie watching decisions?

Analytical Approach: Using data engineered features in combination with critics' ratings platforms to create an optimized model.

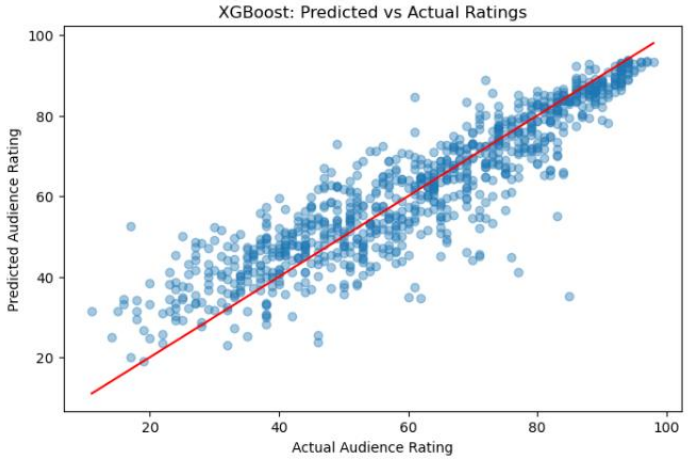
Key Results:

Model	Method	Best Parameter	Best R^2
Linear Regression	5-Fold Cross Validation	N/A	0.747
Polynomial Ridge Regression	Grid Search *degrees: 2,3,4* *alpha:0.1,1,10*	Degree = 2 Alpha = 10	0.632
Decision Tree Regression	Grid Search *max depth: 1 - 6*	Max depth = 4	0.769
Random Forest Regression	Grid Search *max depth: None,1,3,5,7* *estimators: 100 - 1100*	Max depth = none Num Estimators = 700	0.798
KNN Regression	Grid Search *neighbors: 1 - 9*	N Neighbors = 8	0.472
XG Boost Regression	Randomized Search *estimators: 200 - 1600* *learning rate: 0.01 - 0.2* *max depth: 2 - 11* *subsample: 0.6 - 1* *minimum child weight: 1 - 7* *gamma: 0 - 0.3* *lambda: 0,1,5,10* *alpha: 0,0.1,0.5,1* *column sample: 0.6 - 1*	Num Estimators = 300 Learning Rate = 0.03 Max Depth = 7 Subsample = 0.7 Minimum Child Weight = 3 Gamma = 0.1 Lambda = 5 Alpha = 1 Columns Sample = 0.9	0.801
XG Boost Regression	Grid Search *estimators: 250, 300, 350* *learning rate: 0.02, 0.03, 0.04* *max depth: 6, 7, 8* *subsample: 0.6, 0.7* *column sample: 0.8, 0.9*	Num Estimators = 350 Learning Rate = 0.02 Max Depth = 6 Subsample = 0.7 Column Sample = 0.9	0.803

***Model Selection**:* XG Boost achieved the highest average R^2 (0.803) and was selected as the final model.

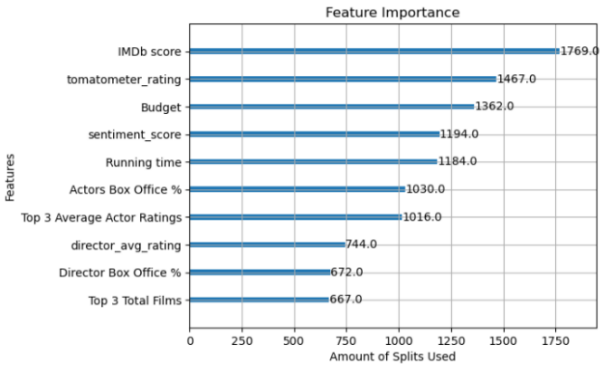
***Model Assumptions**:* The selected XG Boost model uses a low learning rate (0.02) and many estimators (350), allowing errors to be corrected gradually while improving predictive stability. Additionally, tree complexity is controlled through a maximum depth of 6 and a minimum child weight of 3, which helps prevent overfitting and unstable splits. Additional regularization is introduced by using a 70% row sampling for each consecutively built tree and feature subsampling of 90%, increasing model generalization and robustness. Finally, gamma (0.1) and L1/L2 regularization (reg_alpha=1, reg_lambda=5) penalize unnecessary complexity, improving generalization to unseen data.

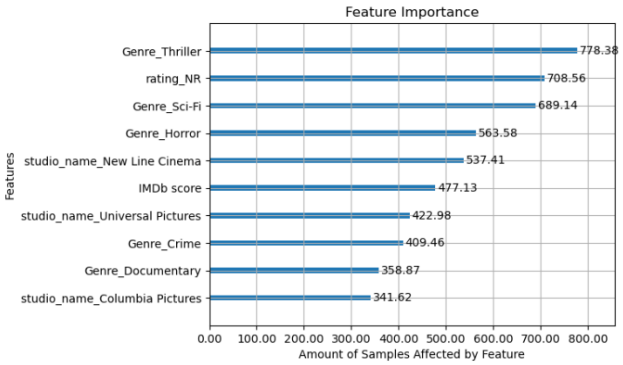
Interpretation: Introducing additional engineered features improved model performance by 3.2%, indicating a meaningful increase in variance explained. This improvement is also evident visually, as predicted ratings align more closely with actual audience ratings, shown by the tighter clustering around the reference line. While a 3.2% increase may appear modest, even small gains are valuable when modeling inherently subjective outcomes such as audience sentiment, providing more reliable insights for both studios and moviegoers.



***Question to Answer**:* Which factors have historically influenced audiences' enjoyment of films?

***Interpretation**:* The feature importance charts below highlight the most important factors (beyond the IMDb Rating) that influence audience enjoyment. The top drivers are the run time of the movie, budget, critics sentiment, genre, and MPAA rating (G, PG, etc.).





6. Interpretation & Managerial Implications

6.1 Interpretation

Our analysis shows that IMDb ratings are the most reliable indicator of whether audiences will enjoy a movie. However, other factors also matter, including a film's budget, runtime, and the overall tone of critical reviews. Movies within certain genres and from consistent studios also tend to align better with audience preferences.

6.2 Managerial Implications

Audience: Choosing what movie to watch can be challenging for audiences. The findings suggest that viewers can make more informed decisions about which movies are worth their time and money by leveraging critic insights strategically. While critic ratings provide a useful starting point, audiences should also consider factors such as runtime, budget, genre, and the overall sentiment of reviews to improve their viewing choices.

Studios: For studios in the film industry, their biggest challenge is predicting how well a film connects with the audience. They can leverage our insights and explore strategic opportunities to enhance audience engagement and revenue. Therefore, they can optimize movie runtime and align release timings with critic sentiment when marketing films.

6.3 Limitations

Our project has a few clear limitations. The biggest risk is overfitting, since the final model went through heavy hyperparameter tuning. With a relatively small dataset, the model's performance may not generalize beyond the sample we used. Another limitation is that we focused primarily on a single modeling approach to XGBoost, because it consistently

performed the best. Although effective, it limits the scope of comparison and may overlook patterns that alternative modeling frameworks (e.g. GAMs) could capture.

7. Conclusion

Our analysis showed that audience ratings are shaped by a mix of numeric and categorical features, and many of these relationships were clearly non-linear. The dataset also highlighted patterns we didn't initially expect, such as how sentiment extracted from critic text consistently aligned with audience ratings once converted into numeric scores. Overall, the data revealed enough variability across films (budget ranges, genres, release timing, etc.) for tree-based models to capture deeper structure than simple linear methods.

Future work could expand the dataset to include more years, more platforms, or audience demographic information. Incorporating marketing spend, franchise indicators, and social media sentiment would also strengthen the model's predictive power. Exploring additional model families such as generalized additive models or neural networks could help capture complex patterns without relying so heavily on optimization. A classification approach (predicting whether a movie will be "well-liked") could also make the results more actionable for real-world applications.

8. Bibliography

Chen, T., Antunovic, S., & Rafatired, S. (2024). *Predicting movie ratings based on metadata*. University of California, Davis. https://www.academia.edu/143507856/Predicting_Movie_Ratings_Based_on_Metadata

Mahmud, Q. I., Shuchi, N. Z., Tawsif, F. M., Mohaimen, A., & Tasnim, A. (2020). *A machine learning approach to predict movie revenue based on pre-released movie metadata*. *Journal of Computer Science*, *16*(6), 749–767. https://doi.org/10.3844/jcssp.2020.749.767