Movie Trend Analysis

Huong Giang Ho
Dept. of Computer & Information Sciences
Towson University
Towson, USA
gho2@students.towson.edu

Ehren Dietrick
Dept. of Computer & Information Sciences
Towson University
Towson, USA
edietri2@students.towson.edu

Aminata Bangura
Dept. of Computer & Information Sciences
Towson University
Towson, USA
abangu6@students.towson.edu

Abstract*— This paper analyzes box office movie data from 2000-2004 to uncover trends, patterns, and predictions within the film industry. Using datasets from Kaggle, data mining techniques will be applied to explore relationships between genre, release dates, budgets, ratings, and revenue. Techniques such as classification, regression, and clustering will be used to identify factors that influence movie performances. The goal is to examine movie trends in the past and create prediction models that can estimate a movie's probability of success based on features from similar movies. Furthermore, this study investigates the viability of a minimal-feature predictive model using movie ratings as a single predictor, benchmarking its performance against a comprehensive multi-variable Random Forest regression model to evaluate the trade-off between simplicity and predictive accuracy.

Keywords*— Movie Analytics, Movie Predictions, Data Mining, Data Analytic

I. Introduction

This project focuses on applying data mining techniques to box office data to uncover meaningful patterns and predictive insights. Our problem statement centers on the question: Can historical movie data be used to identify trends and predict how future movies may perform? By addressing this question, broader relevance of data mining in entertainment analytics will be explored, where uncovering hidden patterns can inform decision-making for filmmakers, studios, and analysts.

The first primary objective of this study is to identify if one variable can have the standalone predictive power of movie ratings as an indicator of box office success, which in contrast to conventional multi-factor models that incorporate variables such as budget, marketing spend, and genre at the same time. The study specifically investigates whether a single feature can reliably predict a film's success across multiple datasets. This focused approach not only clarifies the functional role of public perception in film outcomes but also explores the consistency of rating-performance correlations across different movie dataset. Ultimately, the research aims to contribute insights into the viability of minimal-feature predictive systems in entertainment analytics.

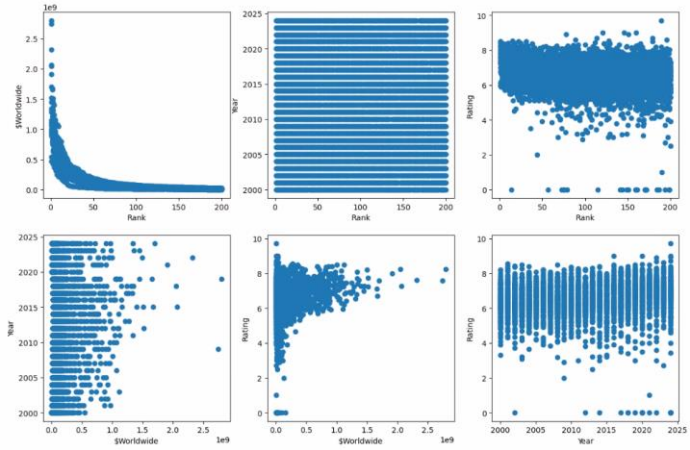
The second primary objective of this study is to benchmark this minimal-feature approach against a comprehensive, multi-variable regression model that utilizes all available predictors in the dataset, including budget, genre, runtime, and release timing. This comparative analysis is designed to quantify the predictive trade-off between parsimony and comprehensiveness, evaluating how much explanatory power is sacrificed or retained when relying on a single proxy like ratings instead of a full suite of conventional factors. By systematically contrasting the performance of the simplified model with that of the holistic model, the research seeks to determine the practical cost of simplicity in predictive accuracy, thereby situating the utility of a single-feature system within the broader context of established analytical frameworks in film industry forecasting.

II. Literature Review

Studies have shown that data mining methods can reveal patterns of movie success factors. For example, budget can predict revenue, while genres can significantly impact performance. The methodologies used in these studies include the use of regression analysis and machine learning, providing a framework for our analysis [4], [5], [6].

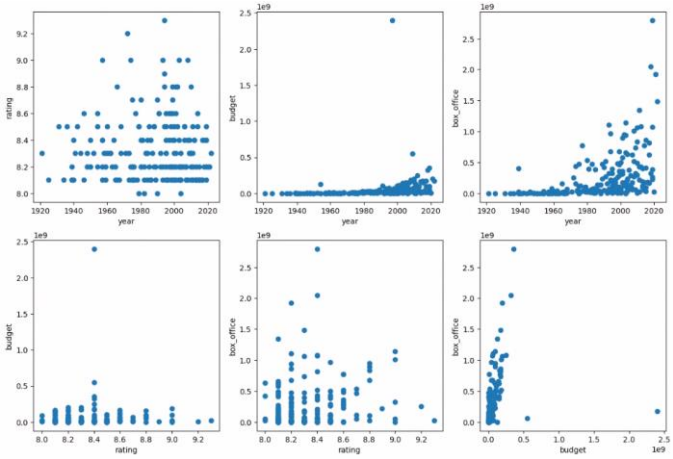
III. EDA

A. Movies Box Office Dataset (2000-2024)



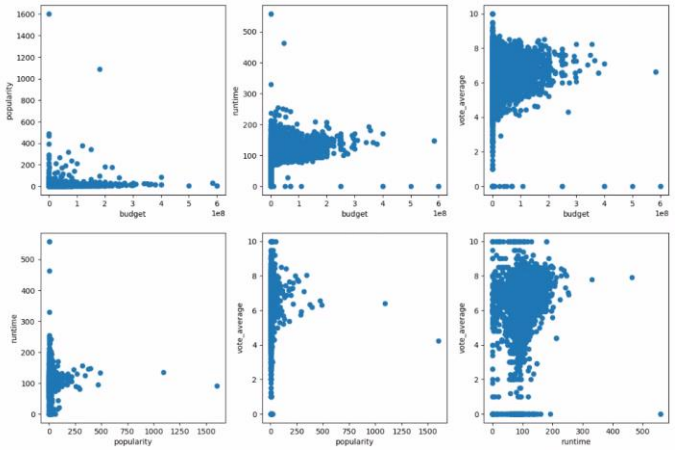
The visual analysis of the Popular Movies dataset reveals several critical relationships between rank, revenue, and rating. First, a steep inverse curve in the Rank vs. Worldwide Revenue plot confirms that the highest-ranked films are predominantly those with the greatest commercial success, with revenue declining sharply for lower-ranked titles. However, the relationship between Rank and Rating displays high variance, indicating that a film's popularity-based ranking does not strongly correspond to its critical reception, as movies across all rank positions achieve a broad range of ratings. This is further illustrated in the Revenue vs. Rating plot, which shows only a slight upward trend; while higher-rated films tend to earn more, the widespread demonstrates that significant revenue can be generated by lower-rated films, and vice versa. Finally, plots of Year against both Rating and Revenue show broad scattering with no clear temporal pattern, suggesting that the stability of ratings and the potential for high earnings have remained relatively constant across the years in this dataset.

B. IMDB Top 250 Movies Dataset



The analysis of the IMDB Top 250 dataset reveals a distinct set of patterns. First, the relationship between Rating and Year shows no discernible trend, as ratings are tightly clustered between approximately 8 and 9.3 across all years. This distribution confirms that the inherent selection bias of a "top" list flattens any temporal effect, isolating only the highest-rated films regardless of release date. Second, examining Rating against both Budget and Box Office demonstrates a critical insight: films with similar, high ratings achieve vastly different commercial outcomes. The presence of dense vertical bands in these plots visually confirms a weak correlation, where mid-high ratings can correspond to both exceptional and modest box office returns. Finally, while the Budget vs. Revenue plot shows a general positive trend, the substantial spread around this trend underscores that significant financial success within this elite group is driven by factors well beyond a film's critical rating alone, highlighting the complex, multi-variable nature of blockbuster performance.

C. Popular Movies Dataset



Based on visual analysis of bivariate relationships, several key patterns emerge. First, plots of popularity versus budget and revenue reveal significant variance, indicating that while high-budget films can achieve substantial popularity, a notable number of modestly budgeted films also reach high popularity. This suggests that commercial success is not strictly tethered to budget or rating alone but is likely influenced by other attributes such as genre novelty, cast, or franchise appeal. Second, the relationships between runtime and both popularity and rating show a broad spread, indicating no strict correlation, as highly popular films exist across the entire spectrum of film lengths. Most notably, the plot of popularity versus rating shows the strongest trend, with lower-rated films clustering at lower popularity scores and higher ratings correlating with a clear, though dispersed, upward trend in popularity. This indicates that while higher ratings are generally associated with greater popularity, the relationship is not deterministic, further underscoring the multifaceted drivers of a film's success.

IV. METHODOLOGY

A. Data collection

Three public datasets on Kaggle: “Movies Box Office Dataset (2000-2024)”, “IMDB Top 250 Movies Dataset”, “Popular Movies Dataset”. These datasets provide

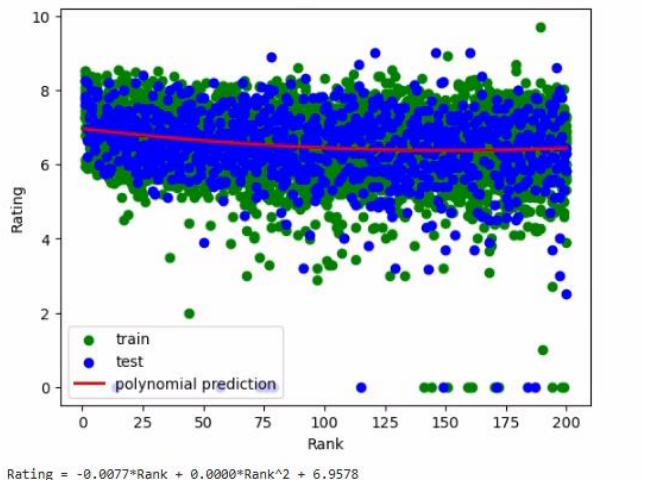
variables such as genres, movie titles, release dates, budgets, and ratings [1], [2], [3].

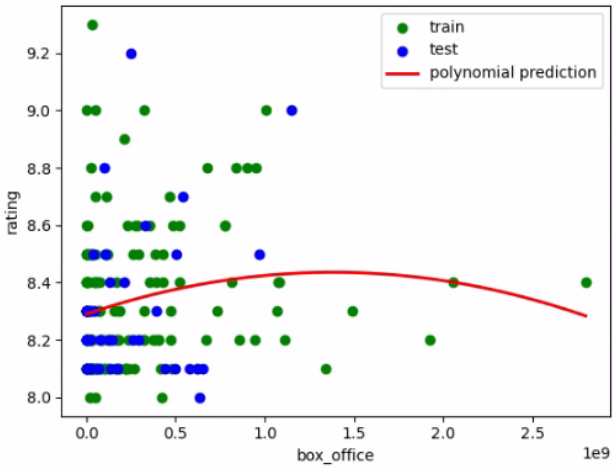
B. Data Preprocessing

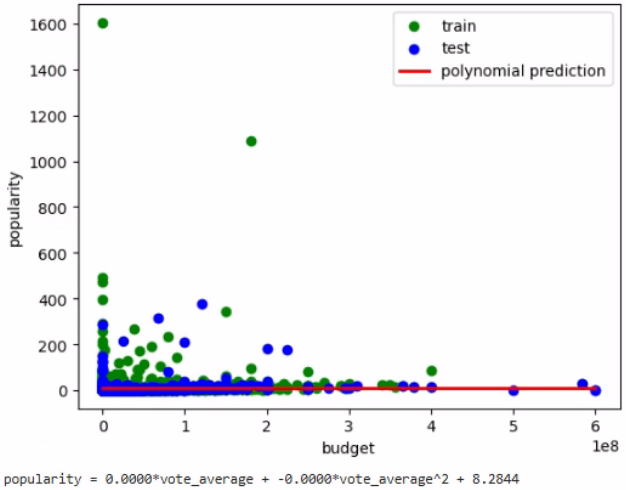
All datasets were preprocessed to ensure data quality and consistency. Removal of duplicate movie records and handling missing or invalid values were involved. Movies with missing ratings, box office revenue, budget or ranking information were excluded. After preprocessing, each dataset was structured into independent variables and a dependent variable for regression modeling.

C. Linear regression

Linear regression was used to evaluate whether a single feature such as movie rating, budget or rank could independently predict box office performance. This aligns with this project's objective of assessing minimal feature predictive models rather than complex multi-variable systems. Separate regression models were created for each dataset to maintain consistency and enable comparative analysis.







D. Comparative Random Forest Regression Model

Random Forest Regression is an ensemble learning algorithm chosen for its robustness in handling complex, non-linear relationships and mitigating overfitting through the averaging of multiple decision trees. To ensure a rigorous comparative analysis, a consistent methodological pipeline was applied across all models. This process began with the independent preparation of the two primary datasets (enhanced_box_office_data and popular_movies), involving the encoding of categorical variables and standardization of all features. The target variable for prediction remained the film's rating (vote_average) across all experiments. Each model configuration whether utilizing the full feature set or a single variable was trained and validated using an 80/20 train-test split, with performance evaluated against the standardized metrics of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) to facilitate direct and meaningful comparison.

V. EXPERIEMENTS AND RESULTS

		Dataset 1	Dataset 2	Dataset 3
Linear regression	MSE	1.1629	0.062	0.0692
	RMSE	1.878	0.23	0.24
	R^2	0.018	0.036	0.037
Random Forest Regression	MAE	0.5008	0.162	0.401
	RMSE	0.6775	0.2066	0.6527
	R^2	0.4435	0.2097	0.7846

A. *Linear regression*

After conducting EDA on each dataset, Rank, box-office and budget all seemed the most likely to be significant. A linear regression was made for each model, and each was found to be statistically insignificant (An r^2 value less than 0.05). With linear regression, with each of our datasets, there seemed to be almost no significance. Each test had an r^2 value below 0.05, meaning the models could not conclude there to be a significant relationship.

B. *Comparation of linear regression and random forest regression*

The Random Forest models demonstrated substantially better performance across all datasets compared to simple linear regression. Dataset 3 showed the strongest results with an R^2 of 0.7846, indicating that the multi-variable ensemble approach could explain approximately 78% of the variance in movie ratings. Interestingly, the single-feature linear regression models performed particularly poorly on Dataset 1 ($R^2 = 0.018$), suggesting minimal linear relationship between any single predictor and the target variable in that dataset. The Random Forest's superior performance highlights the importance of capturing non-linear interactions and feature combinations when predicting movie success metrics.

VI. Discussion

This analysis concludes that while movie rating can inform predictions of success, its reliability is not universal but contingent on the dataset's inherent variability. Models proved effective only when datasets contained a broad spectrum of movie quality and commercial performance. For instance, Dataset 2, with its uniformly high ratings and minimal variance, demonstrated a sharp failure of single-feature prediction, as the near-constant rating offered no discriminative power. Conversely, in datasets like Dataset 3, where ratings exhibited meaningful variance and correlated positively with popularity, even simple models could capture a real predictive trend. Therefore, a rating utility as a standalone predictor is viable not in curated, elite sets, but specifically in non-curated, mixed-quality movie pools where audience reception displays significant volatility.

The superior performance of Random Forest regression across all datasets underscores the complex, multi-factorial nature of movie success prediction. While single-feature models offer simplicity and interpretability, they sacrifice substantial predictive power. The 0.7846 R^2 achieved by Random Forest on Dataset 3 represents a significant improvement over the corresponding linear regression model, demonstrating that comprehensive feature integration provides significantly better explanatory capability. However, this comes at the cost of model complexity and reduced interpretability.

VII. Conclusion

This study demonstrates a clear trade-off between model simplicity and predictive accuracy in movie success forecasting. While single-feature linear regression models offer interpretability and computational efficiency, they fail to capture the complex, multi-dimensional relationships that drive movie performance. The Random Forest regression approach, incorporating multiple features and their interactions, achieved significantly higher predictive accuracy across all tested datasets.

The research confirms that no single feature can consistently provide strong standalone predictive power across diverse movie datasets. Instead, successful prediction requires consideration of multiple factors including budget, genre, runtime, and temporal release patterns. For practical applications in the film industry, we recommend using ensemble methods like Random Forest regression despite their increased complexity, as they provide substantially better predictive performance.

Future work should explore hybrid approaches that balance interpretability with accuracy, potentially through feature selection techniques that identify the minimal set of features needed for near-optimal prediction. Additionally, incorporating more sophisticated features such as social media sentiment, cast popularity metrics, and franchise affiliation could further improve predictive models for movie success.

VIII. Bibliography

[1]

ADITYA JILLA, "Movies Box office Dataset (2000-2024)," *Kaggle.com*, 2024, doi: https://doi.org/10347532/118ff04a5fb9db66a72e1e4cf797a13f.

[2]

"IMDB Top 250 Movies Dataset," *www.kaggle.com*. https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset

[3]

Rajan, "Popular movies dataset," *Kaggle.com*, 2025. https://www.kaggle.com/datasets/rajansavaliya22/popular-movies-dataset?select=popular_movies+%282%29.csv (accessed Nov. 24, 2025).

[4]

J. Lee, "Exploratory Data Analysis With Movies," *Medium*, Sep. 06, 2020. https://medium.com/data-science/exploratory-data-analysis-with-movies-3f32a4c3f2f3 (accessed Nov. 24, 2025).

[5]

H. Yang, Y. Pei, and Z. Wang, "Movie Data Analysis and a Recommendation Model," *Proceedings of the 2022 International Conference on Bigdata Blockchain and Economy Management (ICBBEM 2022)*, pp. 739–751, Dec. 2022, doi: https://doi.org/10.2991/978-94-6463-030-5_74.

[6]

L. Kharb, D. Chahal, and Vagisha, “Forecasting Movie Rating Through Data Analytics,” *Data Science and Analytics*, pp. 249–257, 2020, doi: https://doi.org/10.1007/978-981-15-5830-6_21.

IX. Contributions

Aminata: Provided EDA analysis with histograms and completed methodology for the paper (part III), initiating research topics,

Ehren: Performing some of the EDA analysis, linear regression models of coding, add to the paper, format the data, finding out more about each dataset. (IV B, C; V A)

Huong Giang: Write introduction (I), literature review (II), EDA written analysis (III), Random Forest Regression coding and written analysis (IV D, V B), discuss, conclude, format and finalize paper (VI, VII, VIII, abstract, bibliography)