Predicting Movie Box Office Success: Executive Summary

Erdös Institute Data Science Boot Camp, Spring 2025

Team: Tawfiq Abdullah, Divyantha Malimboda Gamage, Nikhil Nagabandi, Obinna Ukogu
Project Mentor: Collin Litterell

Introduction: Studios must decide early whether to finance a film. Because movie ticket revenues are shared with movie theaters, a rough heuristic is that a movie needs to make twice as much as its budget to be considered a financial success. By predicting return on investment (ROI)—the ratio of worldwide box office to production budget—we aim to offer quantitative guidance to movie studios and film-makers.

Data Collection, Cleaning and Exploratory Analysis: We combined public datasets to collect information on budget, box office revenue (domestic and worldwide), IMDB rating, production company, cast and crew information, and movie synopsis. These were standardized to align records across sources; duplicates and non-theatrical releases were removed. Outliers such as those with budgets below $10k or above $300M, and ones with extreme runtimes were excluded.

All budgets were adjusted for inflation to the value in 2020 US dollars using the average inflation rate over the past 50 years. We one-hot encoded movie genres, and grouped production companies into tiers (major, indie, etc.). For the actors, directors, genres and production companies, we assigned a rolling ‘impact’ score for each year by computing the mean rating/revenue of relevant films in the preceding years. We also computed naive similarity scores between movies based on word frequencies in their synopses.

Models and Results: Our approach to predicting movie box office performance leverages a sophisticated stacked ensemble methodology that combines the strengths of multiple regression algorithms to predict ROI. With root-mean-squared error (RMSE) as our KPI, we evaluated seven regression models using a temporally-aware 5-fold cross-validation framework. This crucial design ensures models only learn from past movies to predict future ones, never the reverse. Maintaining the integrity of our real-world prediction task.

The table below compares the Root Mean Squared Error (RMSE) for individual models and our stacked ensemble:

Model	RMSE (↓ better)	Relative Improvement vs LightGBM
LightGBM	1.9512	Baseline (0 %)
XGBoost	1.8609	+ 4.6 %
CatBoost	1.9556	– 0.2 %
Random Forest	1.9942	– 2.2 %
SVR	2.0264	– 3.9 %
Stacked Ensemble	**1.7573**	**+ 9.9 %**

These results demonstrate that our stacked ensemble delivers substantial improvement over even the best individual model (CatBoost), reducing prediction error by 19.5% compared to our baseline. This confirms that intelligently combining diverse modeling approaches yields superior performance for complex prediction tasks like movie box office forecasting.

Conclusions, Limitations, and Future Directions: While our system shows some predictive power, several challenges remain. First, our dataset was not comprehensive and some independent films are absent from our dataset. Secondly, the accuracy of budget information was hard to verify as studios sometimes understate expenditures on advertising. Another challenge was incorporating information about the content/script of the movie. A more sophisticated model might implement advanced NLP pipelines.