Can we cheat at film making?

A tour into which **IMDb** *features effect a movies community rating and box office*

1

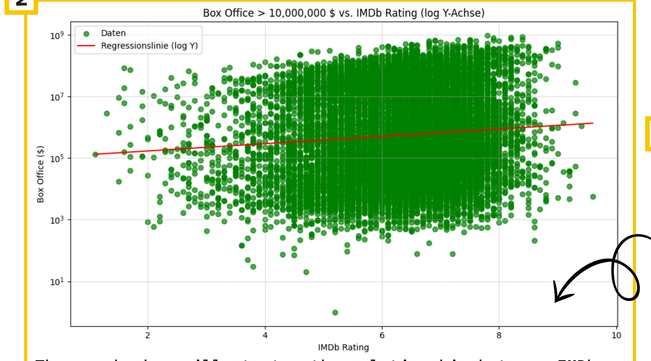
Let's face it – filmmaking is expensive, risky, and emotionally draining. So naturally, we wondered: can we cheat? More precisely, can we reverse-engineer the movie-making process using data and avoid all that pesky "artistic vision" stuff? Inspired by this question (and a deep fear of creative failure), we took a data-driven dive into which IMDb features are actually predictive of a movie's community rating and box office success.

Now, IMDb is many things: iconic, comprehensive, and – as we discovered – pricey. After seeing their commercial dataset pricing, we opted for their free dump to fetch just the IMDb-IDs, then immediately turned to OMDb for the heavy lifting. OMDb: like IMDb, but less expensive and significantly less judgy.

In this poster, we explore what metadata features (cast size, genres, runtime, budget brackets, release years, etc.) move the needle on ratings and revenue. Are there patterns to exploit? Tropes to chase? Hollywood myths to debunk? Possibly. If not, at least we got to complain about IMDb's paywall.

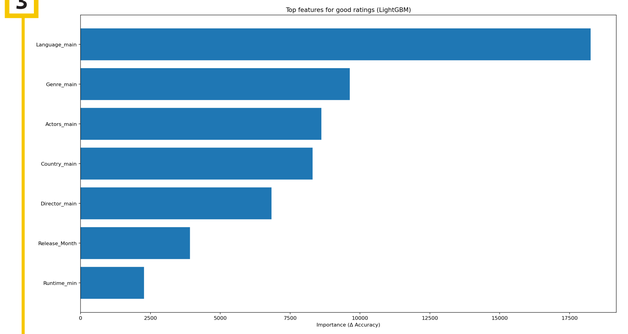
Once we had the data, we decided to split up and each dive into a separate aspect and subset of features.

2



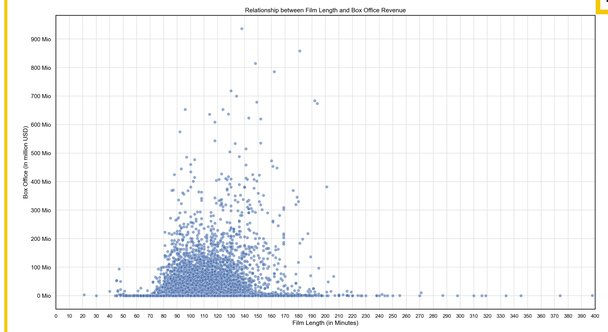
The graph above illustrates the relationship between IMDb ratings and box office revenue for films that earned more than $10 million. Each green dot corresponds to a single film.
The red line represents a linear regression based on the logarithmic box office values. It indicates a slight positive correlation: films with higher IMDb ratings tend to achieve higher box office earnings. However, the data also shows a wide spread, suggesting that other factors play a significant role in a film's financial success.

3



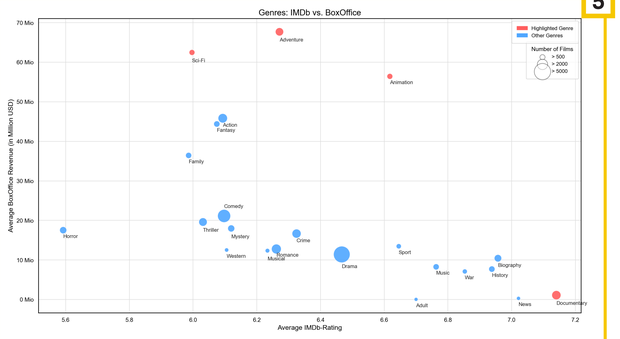
To gauge critical success we labelled every title with an IMDb score ≥ 7 as 1 = hit and the rest as 0 = flop. After basic cleaning the dataset still contained tens of thousands of films, which we split 80 / 20 and fed into a LightGBM classifier. Categorical variables with many levels (director, actors) were target-encoded, low-cardinality ones were ordinal-encoded; numeric fields passed through unchanged. The bars show the permutation loss in accuracy when each feature is shuffled. Primary language dominates: choosing the "right" language (essentially English in this sample) costs the model far more accuracy than any other single tweak, underscoring how strongly global audiences reward linguistic accessibility. Genre comes next – dramas, documentaries and thrillers generally outshine niche horror or slap-stick. Star power (main actors) and production country follow closely; big-name directors matter, but less than expected. Release month and runtime barely move the needle, suggesting timing and length are secondary once content and language resonate.

4



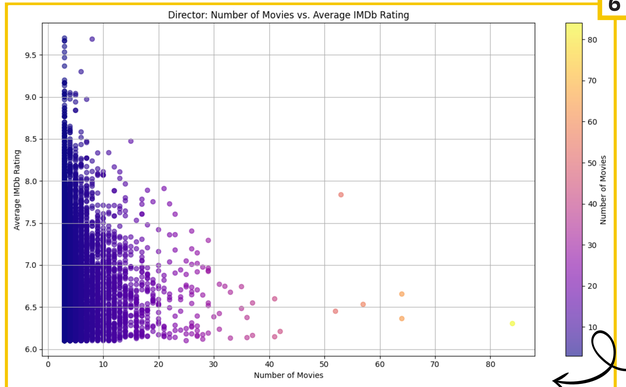
A clear trend emerges from the relationship between film length and box office revenue: the vast majority of commercially successful films fall within a runtime of 90 to 130 minutes. This duration reflects not only the standard format for most feature films but also appears to meet audience preferences. Within this window, many films reach moderate to high box office earnings, suggesting a strong alignment between runtime and market performance. In contrast, films significantly shorter or longer than this range are less common and rarely achieve comparable financial success.

5



Looking at the relationship between average IMDb ratings and box office performance across film genres, a clear pattern emerges, certain genres manage to achieve both critical and commercial success. Adventure, Sci-Fi, and Animation stand out as the dominant blockbuster categories, combining solid audience ratings with the highest average box office revenues. On the other end of the spectrum, Documentary films receive the highest critical acclaim, boasting the highest average IMDb rating above 7, yet generate minimal box office revenue. This genre exemplifies how critical praise does not always translate into financial return. Genres like Drama and Comedy appear frequently in the dataset, as indicated by the size of their bubbles. While they are widely produced, their average ratings and revenues remain moderate. Horror, meanwhile, tends to underperform both critically and commercially.

6



The graph shows the relationship between the number of movies directed by an individual and their average IMDb rating. Each point represents a director.
The visualization reveals that there are directors who consistently perform much better than average.

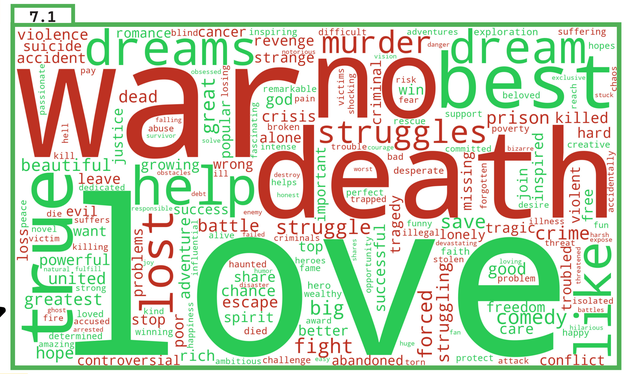
7

Now we point our attention towards the „Plot" field of our dataset. There, the films story is outlined in a few sentences ranging from a minimum of X to a maximum of Y words. According to the IMDb guidelines, the text is sourced from user submissions. There, writing style and rules are also specified. The plot data from OMDb seems to match the IMDb values. For these wordclouds, we have separated the movies into the upper (green border) and lower (red border) .25 percentile to outline potential difference in the plots. We then used a open source sentiment dataset from the Text-Machine-Lab at UMass Lowell to color the words. In addition, we also trained models using RandomForest and XGBoost with varying text vectorizers, to try to predict the rating for a given Plot.

Model	RMSE (lower is better)	R^2 (higher is better)
Dummy (Average)	1.43	0
RandomForest	0.99	0.13
XGBoost	1.01	0.09

This shows that there is at least some information to be extracted from a movies plot

7.1



7.2

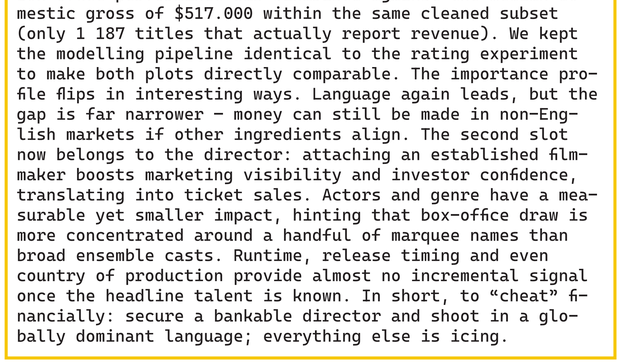
8

Conclusion

In conclusion, our data-driven exploration into the mechanics of movie success suggests that while creativity remains invaluable, certain patterns and predictors can indeed be reverse-engineered. Through metadata analysis and machine learning models, we identified clear signals for both critical acclaim and box office profitability. Language emerges as the single most influential factor across both domains—films in globally dominant languages, especially English, outperform others by a wide margin. For critical success, genre matters: dramas, thrillers, and documentaries are favored, while horror and slapstick tend to underwhelm. Meanwhile, financial success hinges more on marketable elements: the presence of a recognizable director and, to a lesser extent, a notable cast significantly boosts a film's commercial prospects.

Interestingly, factors often cited in industry lore—such as runtime, release date, and country of production—show limited predictive value once higher-impact variables are accounted for. That said, optimal runtimes (90–130 minutes) do seem to align with audience preferences and market viability. Genre analysis reaffirms the dual reality of filmmaking: some categories like Sci-Fi and Adventure manage to be both box office juggernauts and crowd-pleasers, while others, such as Documentaries, thrive critically but struggle commercially.

Even the most subjective narrative elements—like plot summaries—show early promise for differentiating hits from flops using sentiment-based visualization. While we can't yet fully "cheat" the filmmaking process, our findings suggest that strategic data usage can mitigate risk and inform smarter decisions, potentially replacing gut instinct with grounded insight.



Box-office performance was binarised against the median domestic gross of $517.000 within the same cleaned subset (only 1 187 titles that actually report revenue). We kept the modelling pipeline identical to the rating experiment to make both plots directly comparable. The importance profile flips in interesting ways. Language again leads, but the gap is far narrower – money can still be made in non-English markets if other ingredients align. The second slot now belongs to the director: attaching an established filmmaker boosts marketing visibility and investor confidence, translating into ticket sales. Actors and genre have a measurable yet smaller impact, hinting that box-office draw is more concentrated around a handful of marquee names than broad ensemble casts. Runtime, release timing and even country of production provide almost no incremental signal once the headline talent is known. In short, to "cheat" financially: secure a bankable director and shoot in a globally dominant language; everything else is icing.