Explaining Box Office with Minimal Information

Can a Movie’s Success Be Explained with Only Basic Metadata?

Isabelle Huang

December 1, 2025

Movie box office performance is a central topic in both industry and academic research, as it guides investment decisions, marketing strategies, and our understanding of audience behavior. We investigate to what extent simple movie attributes can explain box office outcomes. Using a dataset of 6897 films released between 2006 and 2015, we fit linear regression and shallow neural network models to predict tickets sold from only genre, distributor, MPAA rating, release timing, and a few title-based indicators. Even after tuning, our best linear model attains an R^2 of about 0.35 and the neural network does not substantially improve this, indicating that basic metadata alone are insufficient for accurate box office prediction and that richer information such as budget, marketing, franchise status, and word-of-mouth is likely essential.

Table of contents

1 Introduction

Movie theatres are a common destination for families and friends to gather, enjoy a visual adventure, and briefly escape from real life at a relatively low cost. Yet thousands of films are released worldwide each year. What determines which movies audiences choose, and which titles succeed at the box office? Prior studies have examined determinants such as budget, distribution, sequels, star power, reviews, and word-of-mouth, and have used these variables to model or forecast revenues. In this project, we adopt a similar perspective but ask a more restrictive question: how much of a movie's box office can be explained using only very simple, easily available information?

To investigate this, we fit a linear regression model using only Genre, Distributor, MPAA rating, release date (year and month), and a few indicator variables capturing whether the movie title contains certain frequent words, to predict the number of tickets sold. We then use the linear model's insights to build a shallow neural network and assess whether a more flexible model can substantially improve predictive accuracy using the same limited feature set. This allows us to test whether movie box office outcomes can be adequately captured by basic categorical attributes and coarse timing information alone.

Our results suggest that these simple variables are far from sufficient. Even after tuning, the linear regression model achieves an R^2 of only about 0.35, indicating that most of the variation in ticket sales remains unexplained. The neural network, trained on the same inputs, does not meaningfully improve predictive performance. Among the predictors we consider, Genre has the largest impact on tickets sold, suggesting that audiences exhibit clear genre preferences when deciding what to watch. However, the overall low explanatory power of our models is consistent with prior work showing that box office performance also depends on factors we do not observe here, such as production budget, distribution scale, franchise status, star power, and word-of-mouth dynamics.

Although our models are not accurate enough to be useful for practical forecasting, they highlight an important point for both researchers and practitioners: movie success cannot be captured by a handful of simple metadata fields. For investors, this underscores the value of obtaining richer information about a project—such as budget, marketing plans, and audience buzz—before making decisions. For future research, our findings motivate incorporating additional predictors (e.g., budgets, star metrics, review scores, social media indicators, and country-specific variables) and exploring more sophisticated modelling frameworks.

The remainder of this paper is structured as follows. Section 2 describes the dataset and the cleaning procedures. Section 3 presents our modelling approach and empirical results. Section 4 discusses the findings and limitations, and outlines directions for future work.

2 Data

2.1 Overview

We use the statistical programming language Python [@citePython] together with the data library Pandas [@citePandas] to clean and analyze our dataset. The data come from the Data and Story Library (DASL), an open-source repository of real-world datasets for teaching and practice. Our dataset contains 6 897 major films released between 2006 and 2015, with variables including title, genre, distributor, MPAA rating, release date, and gross revenue.

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (**?@fig-bills**), from @palmerpenguins.

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.4 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \alpha + \beta_i + \gamma_i && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\beta &\sim \text{Normal}(0, 2.5) && (4)\\
\gamma &\sim \text{Normal}(0, 2.5) && (5)\\
\sigma &\sim \text{Exponential}(1) && (6)
\end{aligned}
$$

We run the model in R [@citeR] using the `rstanarm` package of @rstanarm. We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

C References