Explaining Box Office with Minimal Information

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

Isabelle Huang

December 14, 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, only few are remembered and or even watched by more than a few audience. 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. Many studies used complicated models with many predictors to accompolish this task, but is there a way to model the same thing with minimal predictors and simple models? If so, we can make our decisions with much less cost, both in collecting data and performing analysis. In this project, we ask this 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 main estimand is the expected number of tickets sold for a movie with given basic attributes. Let Y_i denote the number of tickets sold for film i, and let X_i be the set of features: genre, distributor, MPAA rating, release year, release month, and three indicators for whether the title contains the words "man", "love", or "life". Our target is the regression function

$$m(x) = \mathbb{E}[\, Y_i \mid X_i = x \,]$$

that is, the average ticket sales for all movies. The linear and neural network models we fit are different parametric approximations to this conditional expectation.

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. 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, for example, budgets, star metrics, review scores, 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 Outcome variables

Gross revenue is an aggregate measure that depends on realized attendance, ticket prices, and the length of theatrical run given in units of USD. The tickets_sold variable is an approximate count based on information provided by theatres and production company. Release dates are recorded as calendar dates and we transform them into release_year and release_month to capture broad timing effects such as releases during holidays. As gross revenue can be easily computed from the number of tickets sold, we will focus our analysis on the number of tickets sold. Since this value is numerial in nature, making Linear Regression a natrual choice for our model.

The number of tickets sold is highly skewed, with a small number of blockbusters selling millions of tickets, while many sell only a few thousand or less, making a large clump near zero when visualized in the original scale. The mean number of tickets sold is about 1.77 million, but the median is only 22,962, emphsizing that the average is pulled up by the outliers with many tickets sold. Because of this, standard deviation is also very large at 5.45 million, reflecting the wide variation in ticket sales across films. This may cause issues for our linear regression model, which assumes normally distributed errors.

We can visualize the top movies with the most and least tickets sold in the following graphs.

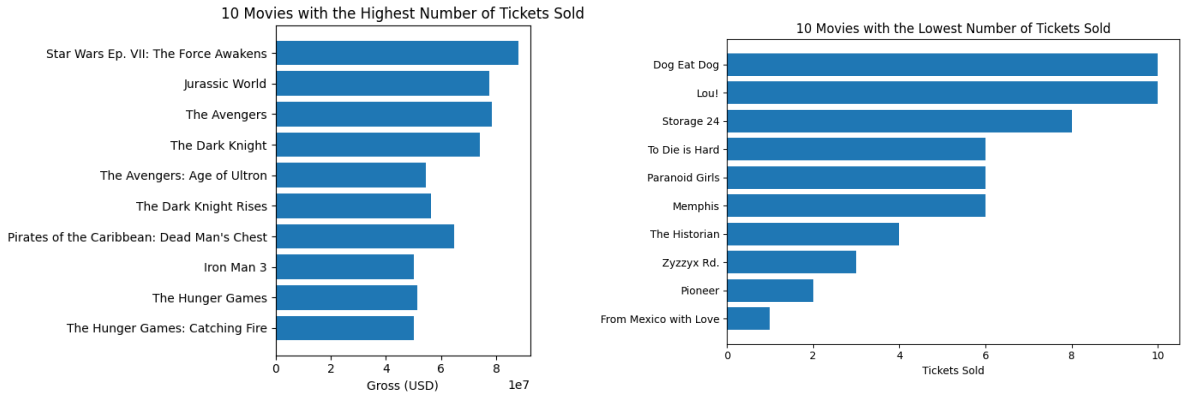

Figure 1: Ticket sales for the ten best- and worst-selling movies.

Because of the clump near zero, we will work with log-transformed version of tickets sold, taken with log base 10. This transformation compresses the scale and reduces the influence of extreme values, making the distribution more symmetric and closer to normal. The mean of log_tickets_sold is about 7.38, the median is about 10.0, and the standard deviation is about 2.87. Althuogh we can still see that the transfored distribution is slightly bi-modal, suggesting a potential mixture distirbution.

2.3 Predictor variables

We consider the rest of the variables in the dataset, except gross revenue, as potential predictors for tickets sold. Genre is a categorical variable indicating the primary genre of the film, such as Action, Comedy, Drama, etc. We can see that the most popular genre in the past year were Distributor is also categorical, representing the company responsible for distributing the film to theatres. MPAA rating indicates the content rating assigned by the Motion Picture Association of America, such as G, PG, PG-13, R, etc. Release year and release month are numerical variables capturing when the film was released.

On top of these, we want to make use of the title of the movies by extracting features from them. We wonder if includig certain key words in the title would increase the chance of the

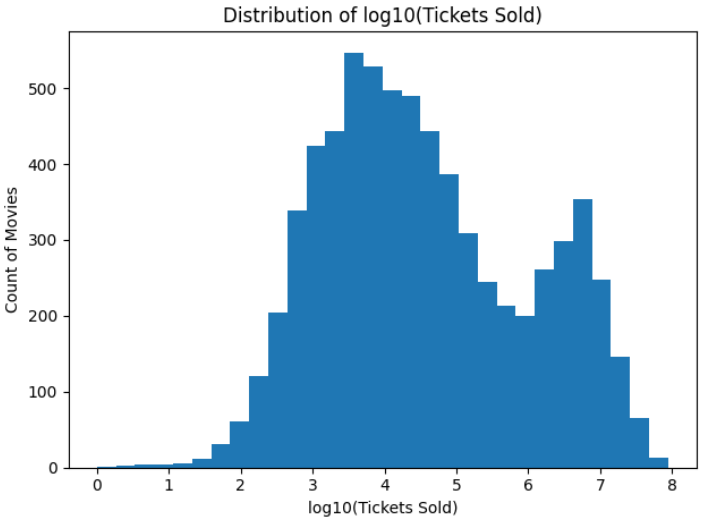

Figure 2: Number of tickets sold on the log base 10 scale

movies being selected by potential audience, thereby increasing the number of tickets sold. So we extracted the most frequent words in appearing in movie titles, which are "man", "love", and "life", and created three indicator variables for each of them. That is, each of these three new variables takes value 1 if the corresponding word appears in the title, and 0 otherwise. These three new variables were aggregated to the dataset, named: title_has_man, title_has_love, title_has_life.

Otherwise, the dataset was well-curated and no more cleaning were required. But it is worth noting that this is an observational dataset of commercially released films, measurement error may be present—for example in ticket count approximations or genre assignments—but we treat the recorded values as given.

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.

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

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