W203 Lab 02 Report - Team 02

Ayoade Israel, Divya Menghani, Hamsini Sankaran, Sivakumar Thiyagarajan

1 Introduction

The movie industry is a multi-billion dollar business, and the budget of a film has become a crucial factor in determining its success, particularly in terms of revenue. Movie production firms, who have control over the budget, are highly motivated to enhance their movie revenue. In 2021, the worldwide revenue of the movies and entertainment market was estimated to be USD 90.92 billion. It is projected to grow at a compound annual growth rate (CAGR) of 7.2% from 2022 to 2030, and by the year 2030, it is anticipated to reach USD 169.68 billion[1].

While there have been several research going on with revenue prediction of movies, over the years, there have been instances where movies with large budgets and compelling plotlines failed to perform at the box office, such as Mulan (Budget: $200 million, Revenue: $69 million), while movies with relatively lower budgets succeeded, such as Super Size Me (Budget: $65K, $22 million)[2]. It raises questions about what factors influence a movie's success.

Our research aims to gain insights into the factors that affects movie revenue and provide recommendations for filmmakers to optimize their marketing and production strategies, using a comprehensive dataset. One of the drawbacks in existing research in predicting movie revenue is the difficulty in accurately capturing the impact of other complex factors that influence a movie's success apart from budget. In this study, by applying a set of regression models, we estimate the revenue of a movie based on budget along with several other factors.

2 Data and Methodology

We gathered the dataset from Kaggle. The dataset comprises metadata of more than 700,000 movies listed in the TMDB Dataset. It has 722986 rows and 20 columns representing various factors influencing the movie's success. It is relevant and provides an opportunity to analyze the relationship between the budget and the revenue of the movies. The dataset contains unique information on movies, with each movie being distinct from the others and identified by an id column in the dataset.

We performed exploratory studies and model building on a 30% subsample of the data. The remaining 70%, totaling 9832 rows, is used for testing the model and evaluation the results in this report. To meet the large sample assumptions, we are evaluating the assumptions of I.I.D and Unique BLP. From the exploratory data analysis, although the distribution of budget and revenue was roughly identical, we observed that our dataset is not strictly I.I.D (further discussed in Limitations section) and it is not having finite variance from the histogram plots, eventually violating the unique BLP assumption. To overcome this, we performed log transformation of the predictor variables (budget and title length). Because each movie is not subjected to multiple samplings, the dataset is regarded as cross-sectional.

We performed data cleaning by ignoring the rows with missing values in all of the dependent and independent variables. From the exploratory plots, we observed few outliers with respect to budget and we are considering the movies with budget greater than $999 and excluding movies with budget greater than $799 million. After filtering the data, we are left with a sample size of 9832 observations.

To best fit the relationship observed in exploratory plots, we operationalize the variables Budget(X) and Revenue(Y) in terms of amount (in dollars, $). Figure 2 shows plot of revenue as a function of budget,

[1] Grand View Research, "Movies And Entertainment Market Size, Share & Trends Analysis Report, By Region, And Segment Forecasts, 2022 - 2030".

[2] BuzzFeed, "10 Films That Flopped At The Box Office Despite Mammoth Budgets, And 9 Low-Budget Movies That Made Hefty Profits", 2023.

depicting a linear relationship between the two variables. Figure 1 shows the person correlation between dependent and all independent variables. Based on the correlation score, we chose the following covariates that affect the dependent variable (revenue), and by including them in our analysis, we will get a better understanding of the fundamental factors that influence the movie revenue.

- vote_count (count): An indicator of the level of engagement a movie has received from the public.
- runtime (minutes): An important aspect of the movie-watching experience that has a correlation with revenue.

Figure 1. Correlation Plot

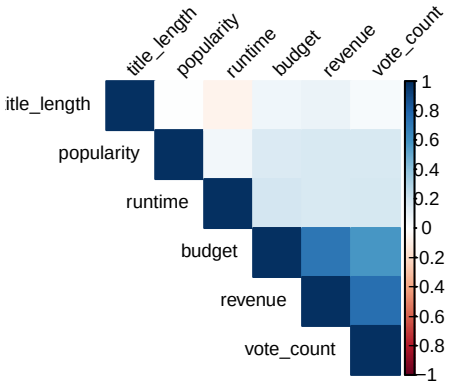
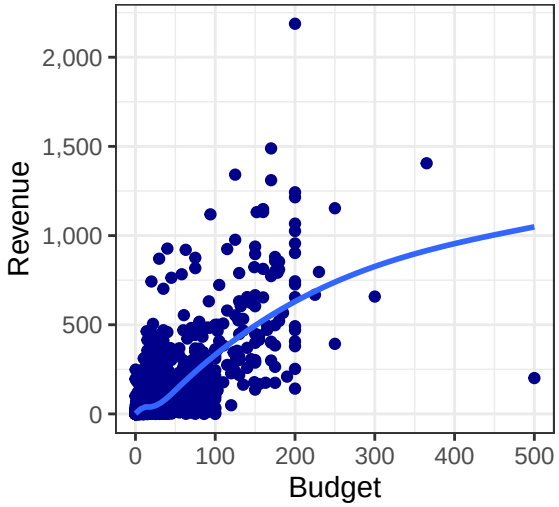


Figure 2. Budget vs Revenue



- popularity (score): A comprehensive metric that takes into account various factors such as promotions, and social media mentions.
- title_length (count): A factor that can impact audience interest, with shorter titles being easier to remember and share. This column is derived from the 'title' column in the dataset.
- release_season (binary): Many film studios release their biggest and most anticipated movies during during Christmas and New Year to take advantage of the increased audience turnout and box office revenue potential. Here we considered two prominent holiday months (December & January) as one category and rest of the months as a different category.
- release_language (binary): Movies released in widely spoken languages such as English have a higher potential audience and therefore may generate higher revenue. By including release language as a covariate, the model can capture this effect and make more accurate revenue predictions. Here we took English as a primary category and all other languages as a non-English category.

Before running the regression model, we checked the mutli-collinearity between all predictors by running a variation inflation factor to reduce unstable and unreliable estimates of the regression analysis and observed that there was no mutli-collinearity between the predictor variables.

We created three regression models on the 30% dataset to understand better to understand the factors contributing to the revenue of a movie. The first model studied the correlation between budget and revenue. In the second model, we gradually added covariates such as 'Vote Count' and 'Run time'. Finally, in the third model, we added more covariates such as Popularity, Movie Title Length, Release season, and Release Language.

$$\widehat{revenue} = \beta_0 + \beta_1 \cdot budget + \mathbf{Z}\gamma$$

where β_0 is the constant in the model, β_1 represents the budget coefficient, $\mathbf{Z}$ is a row vector of additional covariates, and γ is a column vector of coefficients.

Additionally, we evaluated homoscedasticity to check if the variance of the errors or residuals in the model is constant across all the predictor variables, but the results showed heteroscedastic behaviour. Since our dataset is large, we will continue to build our model with robust standard errors to handle the heteroscedastic behaviour.

3 Results

Table 1 shows the results of three regression models with 70% dataset. We can also observe from the Table 1 that Model 3 has the highest adjusted R-squared value of 0.67, indicating that it is the best fitting model out of the three.

Table 1: Estimated Regressions

	Output Variable: Revenue of the movie		
	(1)	(2)	(3)
Budget	0.909^{***}	0.780^{***}	0.772^{***}
	(0.008)	(0.009)	(0.010)
Vote Count		0.0002^{***}	0.0002^{***}
		(0.00001)	(0.00001)
Run time		0.007^{***}	0.008^{***}
		(0.001)	(0.001)
Popularity			0.001^{***}
			(0.0002)
Movie Title Length			0.107^{**}
			(0.036)
Release season			0.181^{***}
			(0.049)
Release Language			0.156^{**}
			(0.051)
Constant	0.514^{***}	-0.325^{***}	-0.800^{***}
	(0.026)	(0.072)	(0.128)
Observations	6,880	6,880	6,880
R^2	0.625	0.671	0.673
Adjusted R^2	0.625	0.671	0.673
Residual Std. Error	1.734 (df = 6878)	1.624 (df = 6876)	1.620 (df = 6872)
Note:	*p<0.1; **p<0.05; ***p<0.01. HC_1 robust standard errors specified in parentheses. Release season is Dec/Jan and rest of the months are considered as non-holiday season. Release Language Category is either English or Non-English movies.		

The co-efficient of the X-concept variable "Budget" is having higher statistical significance in all the three models. Point estimates range from 0.91 to 0.77 which shows that the variable budget has a positive relationship between the other co-variates and the revenue. It is also evident from the table that all the co-variates are statistically significant, having p-value less than 0.05 (alpha level). The co-variates vote count, runtime and release season shows more statistical significance than other co-variates. We also conducted

'Wald' test to study the statistical significance of the two categorical co-variates (Release Season and Release Language) and the results reassured that they are also statistically significant variables in the model.

To better understand the results, consider a hypothetical use case with 180 minutes movie being produced with one million dollar budget in English language with a shorter title length of 10 and planning to release during early December, the Model 3 shows the revenue of the movie could increase to $772,000

From Model 3, we observed that on an average, the model predicts that for every one million dollar increase in the movie budget, the revenue of the movie is expected to increase by approximately $774,800 keeping all the other co-variates constant with an uncertainity of $9574. On an average the model also predicts that for every vote the movie gets, the revenue of the movie is expected to increase by $207.1 (which is a very small effect), keeping all the co-variates constant. Additionally, the model also shows that for every one minute increase in the runtime of the movie, the revenue of the movie is expected to increase by approximately $7,411 keeping all the other co-variates constant. Similarly, the model also shows that for every movie which is released in English language, the revenue of the movie is expected to increase by approximately $156,010 keeping all the other co-variates constant. Likewise, on average, the model predicts that for every one unit increase in the popularity of the movie, the revenue of the movie is expected to increase by approximately $938.

The results emphasizes that budget is a crucial factor in determining the revenue of a movie along with other co-variates vote count, runtime and release season.

4 Limitations

The Dataset has a chance of introducing Sampling Bias as the data about the movies is self reported on TMDB which is a user-generated content platform. Thus, the dataset may represent some movies but dataset cannot be a representative of all movies that have been produced globally. The data may also be biased towards movies that are more popular then the other.

Upon closely examining the movies dataset, the I.I.D. assumption becomes questionable because two movies may share the same cast & crew, production company, and have identical release dates.

The dataset does not mention any information about Motion Picture Association of America(MPAA) Ratings, which have the possibility of influencing the revenue. Likewise, actors may influence both budget & revenue of the movie. The dataset may not represent true picture of current movie industry/trends since it contains information of movies which has a release date before 2013. As the movie industry is rapidly changing, trends become outdated quickly. Moreover, their is no consideration of inflation when concluding the statistical analysis between covariates such as budget & revenue. These variables may introduce omitted variable bias in our model.

Also, in the model we have an outcome variable "Vote count" on the RHS. This variable can be an outcome of the predictor "Popularity". The reason could be because of promotions, more people can engage with the movie which could lead to increase in the number of votes.

5 Conclusion

This study estimated the economic value of movie production, specifically examining the relationship between movie budget and revenue. We also found that several covariates, such as vote_count, runtime, popularity, title length, release season, and release language, have a significant impact on movie revenue. Additionally, we identified the holiday season and language as other important categorical factors that can impact the success of a movie. Our hope is that this line of work will provide filmmakers with accurate tools to plan their investments and optimize their production strategies, reducing uncertainty in the film industry. Future work may explore the correlation of revenue with addition data like MPAA ratings, cast/crew information, sequel information on previous release for identifying the latest trends in the movie industry.