W203 Lab 02 Report - Team 02

Ayoade Israel, Divya Menghani, Hamsini Sankaran, Sivakumar Thiyagarajan

0.1 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. The global movies and entertainment market size was valued at USD 90.92 billion in revenue in 2021 and is expected to expand at a compound annual growth rate (CAGR) of 7.2% from 2022 to 2030 reaching USD 169.68 billion by 2030 [1].

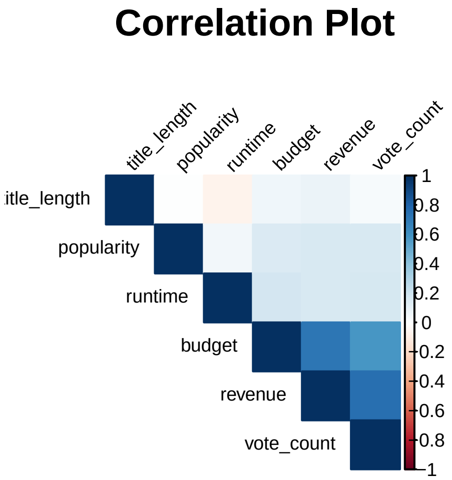
While there have been several research going on with revenue prediction of movies, over the years, there have been instances where movies with large budgets and compelling plotlines failed to perform at the box office, such as The Green Lantern ($220 million), while movies with relatively lower budgets succeeded, such as The Blair Witch Project [2]. It raises questions about what factors influence a movie's success.

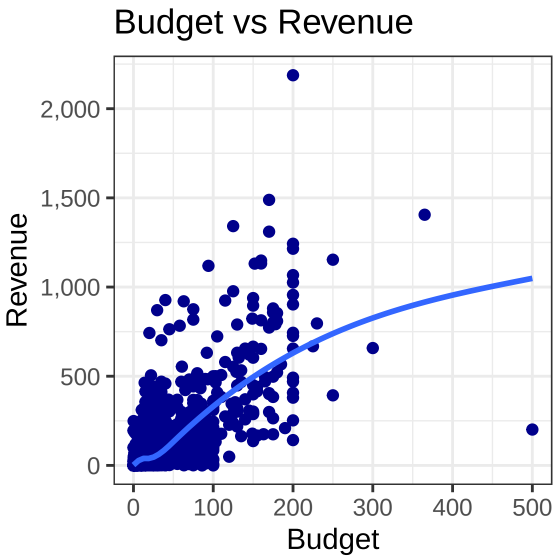
Our research aims to gain insights into the factors that affect movie revenue and provide recommendations for filmmakers to optimize their marketing and production strategies using a comprehensive dataset. One of the drawbacks in existing research in predicting movie revenue is the difficulty in accurately 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.

0.2 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 films, with each movie being distinct from the others and identified by an id column in the dataset. Additionally, the distribution of the X and Y concept is close to identical, meeting the assumptions of IID. Because each movie is not subject to multiple samplings, the dataset is regarded as cross-sectional. Each row in the data represents a movie. 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.

We performed data cleaning by ignoring the rows with missing values in any of the dependent and independent variables. To determine the factors that affects the movie revenue, we performed exploratory data analysis. From the 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 1 shows plot of revenue as a function of budget, depicting a linear relationship between the two variables. Figure 2 shows the person correlation between dependent and all independent variables, from which we observe co-variate having huger correlation with the dependent variable.

From Figure 2, based on the correlation score, We chose the following covariates that has an impact on the dependent variable. By including these covariates in our analysis, we can better understand the key factors driving movie revenue.

- vote_count: An indicator of the level of engagement a movie has received from the public.
- runtime: An important aspect of the movie-watching experience that has a correlation with revenue.
- popularity: A comprehensive metric that takes into account various factors such as promotions, and social media mentions.
- title_length: A factor that can impact audience interest and expectations, with shorter titles being easier to remember and share. This variable is derived from the 'title' column in the dataset.
- release_season: The release season of a movie can impact its revenue. For example, movies getting released during Christmas and New Year tend to have higher revenue. Many film studios release their biggest and most anticipated movies during this time 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: The language in which a movie is released can also impact its revenue. For example, 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. To meet the large sample assumptions, we are evaluating the assumptions of I.I.D and Unique BLP. From the analysis, we observed that our dataset is not strictly I.I.D and not having finite variance from the histogram plots, which violates unique BLP assumption. To overcome this, we performed log transformation of the predictor variables (budget and title length).

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.

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

where β_0 is the constant in the model, β_1 represents the budget coefficient, $\mathbf{Z}$ is a row vector of additional covariates (Vote Count, Run time, Popularity, Movie Title Length, Release season, and Release Language), 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 for the 70% dataset and to handle the heteroscedastic behaviour, we will use robust standard errors.

0.3 Results

Multiple regression analysis are carried out to find the influence of budget over the revenue of a movie.Table 1 shows the results of three regression models. The co-efficient of the X-concept variable "Budget" is having high statistical significance in all the three models studied. Its point estimates ranges from 0.77 to 0.90. We have also conducted Wald test to find out the statistical significance of each co-variates in the model.

0.4 Limitations

After closely observing the movies dataset we can assume that the data is not strictly I.I.D. as two movies can be having the same cast & crew or same production company and exactly same release dates. Also by the scatterplot we can observe a strong co-relation between revenue vs budget and revenue vs runtime, thus we can say that columns in the movie dataset are not independent of each other. 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 not be representative of all movies that have been produced globally. The data may also be biased towards movies that are more popular then the other. The dataset does not mention any data about Motion Picture Association of America(MPAA) Ratings hence dataset does not gives us any information about the anticipated age groups of the viewers. The dataset has lot of NA or zero values for key columns such as revenue, budget and vote_count, after the data cleaning process the size of actual dataset has reduced considerably. The dataset represents some extreme outliers such as $0 value for budget, revenue and runtime of more then 1000 minutes (16.66 hr), these outliers can influence the overall statistical analysis of the dataset. 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, movie industry is a fast paced world and the trends get outdated very soon moresoever their is no consideration of inflation when concluding the statistical analysis between covariates such as budget & revenue.

0.5 Conclusion

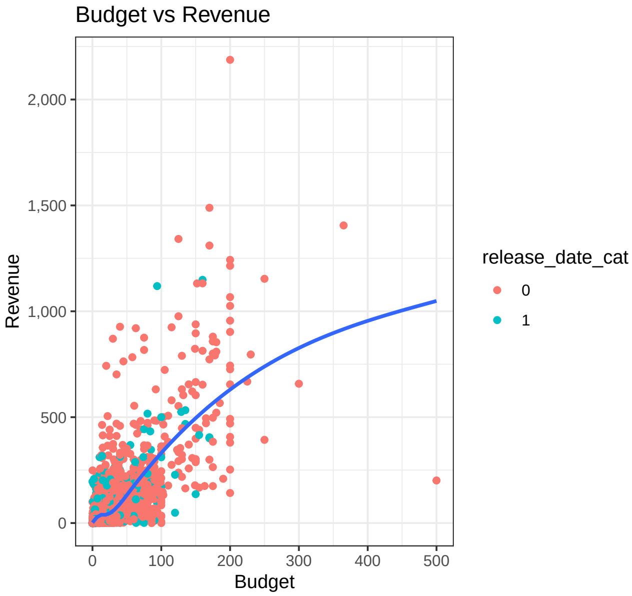
This study estimated the economic value of movie production, specifically examining the relationship between movie budget and revenue.We also found that several covariates, such as vote_count, runtime, popularity, title_length, release_date_cat, and language_cat, have a significant impact on movie revenue. Additionally, we identified the holiday season and language as 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 research could examine the value of specific production decisions, such as marketing strategies or casting choices.

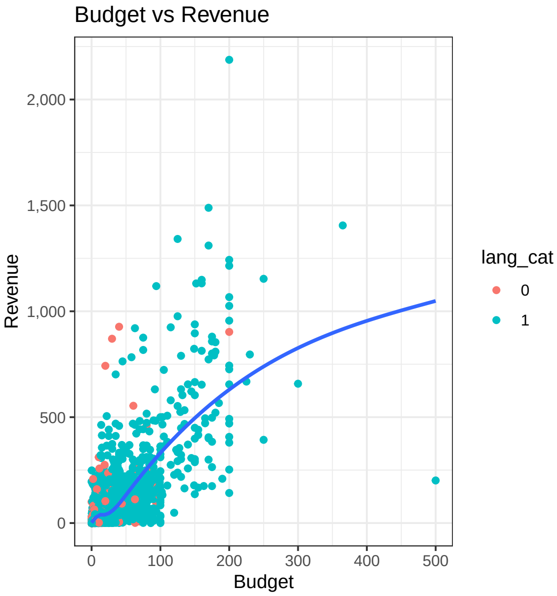
Table 1: Estimated Regressions

	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. Holiday season is Dec/Jan and rest of the months are considered as non-holiday season. Language Category is either English or Non-English movies.

0.6 Appendix


Budget vs Revenue
Revenue
2,000
1,500
1,000
500
0
Budget
0
100
200
300
400
500
release_date_cat
● 0
● 1


Budget vs Revenue
Revenue
2,000
1,500
1,000
500
0
Budget
0
100
200
300
400
500
lang_cat
● 0
● 1