Data Science Project
> Box Office success determinants and future recommendations <

Introduction

After a series of Box Office flops, the studio recognized the need for a change in direction. Using a sample of 5043 movies, we are asked to identify determinants of a successful movie, as well as to make future recommendations. This study's aim is to define key correlations between determinants, explore their relevance, and make recommendations on how they can be used to improve Box Office performances.

The Data

This study used data provided by the studio, with a total number of 5043 movies dating from 1927 to 2016. It contains movies from 65 countries, mainly the USA, in both color and black and white. Data provided information on the movie's revenue (Box Office gross), production cost, main cast, director's information, aspect ratio and content rating, the language spoken, Facebook popularity measured in the number of likes, movie's genre and plot keywords, as well as numbers of user votes, reviews, critic reviews and the movie's score. Missing data was eliminated from the sample, as about 30% of it was located in the movie's gross, and budget data. The sample was further filtered down by removing existing duplicates. The final sample contained a total of 3411 movies.

Target value

Target value used in this study is the Box Office *gross* in US dollars. An assumption was made that the *gross* value is corrected for inflation.

Feature transformations

- ➔ *Genre* feature was filtered down to contain only the top 4 genres based on the assumption these genres were the most popular as they hold the majority of values. These genres are: Comedy (1329 values), Action (1153 values), Adventure (453 values), and Drama (972 values).
- ➔ *Content rating* was filtered down to contain the two most frequent ratings: R and PG-13.

New features

Two new features were created to aid the data analysis process:

- ➔ *Profit* was calculated by subtracting the production cost (in this case budget) by the total amount earned (gross).

- ➔ *Roi* was calculated by subtracting the budget by net profit, and multiplying it by 100. The resulting amount is in %.

Eliminated features

- ➔ *Color* - there is no evidence this feature has any relevance to the target feature.
- ➔ *Aspect Ratio* - there is no evidence this feature has any relevance to the target feature.
- ➔ *Actor 3* and *Actor 3 Facebook Likes* - An assumption was made that additional cast doesn't have any relevance for future recommendations, so the Actor 3 feature was eliminated, along with the corresponding Facebook popularity likes.
- ➔ *Director's Name* and *Director's Facebook Likes* - there is no evidence a director's name might influence profit or gross, along with the corresponding Facebook popularity.
- ➔ *Plot Keywords* - this feature contains over 500 unique values. There is no evident method of finding a pattern in such a vast number of unique values, and if it will aid the exploration process.
- ➔ *Country and Language* - Initial exploration identified huge outliers in the budget feature, which were actually amounts expressed in the currency of the country of origin. Further analysis confirmed there are more instances of a different currency in foreign countries, while English speaking countries had the same currency: US dollar. The country feature was removed, as it served no obvious purpose in the analysis process, while the language feature was filtered down to only English speaking values.

Data Analysis

Box Office Gross

The average mean value for gross is $53.06M, with a median value of $30.98M. Highest grossing movie is ‘Avatar’ with $760.51M, with a budget of $237M and a 220% roi. Lowest grossing movie is ‘Skin Trade’ with $162 gross, $9M budget and a roi of -99.99%. The gap between the mean and the median values is obvious, as the values lie in the 60th percentile.

Average values of the top 12 grossing movies show us an obvious broad range from the entire sample. The average gross mean is $513.86M, ten times the average gross value. This trend is also seen in the production cost which averages to $459.96M. Median values do not stray as much from the mean, as seen in the entire sample. Highest ROI values in the top 12 belong to ‘Star Wars: Episode IV - A New Hope’ with a ROI of 4090.3%, and ‘E.T. the Extra-Terrestrial’ with ROI of 4042.3%. Both titles were released two or three decades prior to the rest of the group. This supports the finding that the mean production cost for movies released before 1990 is $13.87M, while the mean production cost for titles between 1990 and 2016 is $40M, three times higher.

> Table 1: Top 12 highest grossing movies <

No.	Movie Title	Gross (US $ Million)	Budget (US $ million)	ROI	Genre	Year
1.	Avatar	760.51	237.0	220.8%	Action	2009
2.	Titanic	658.67	200.0	229.3%	Drama	1997
3.	Jurassic World	652.18	150.0	334.7%	Action	2015
4.	The Dark Knight	533.32	185.0	188.2%	Action	2008
5.	Star Wars Episode I: The Phantom Menace	474.54	115.0	312.6%	Action	1999
6.	Star Wars Episode IV: A New Hope	460.94	11.0	4090.3%	Action	1977
7.	Avengers: Age of Ultron	458.99	250.0	83.5%	Action	2015
8.	The Dark Knight Rises	448.13	250.0	79.2%	Action	2012
9.	Shrek 2	436.47	150.0	190.9%	Adventure	2004
10.	E.T. - The Extra-Terrestrial	434.95	10.5	4042.3%	Family	1982
11.	The Hunger Games: Catching Fire	424.65	130.0	226.6%	Adventure	2013
12.	Pirates of the Caribbean: Dead Man's Chest	423.03	225.0	88.0%	Action	2006

A simple correlation matrix shows us the values of correlation between gross and other variables:

> Table 2: Correlation matrix <

	Gross	Budget	# users voted	# user reviews	# critic reviews	Movie Facebook likes	Movie Score
Gross	1.00						
Budget	0.63	1.00					
# users voted	0.62	0.40	1.00				
# user reviews	0.54	0.41	0.79	1.00			
# critic reviews	0.45	0.45	0.59	0.56	1.00		
Movie Facebook likes	0.34	0.30	0.52	0.38	0.71	1.00	
Movie Score	0.25	0.06	0.50	0.35	0.37	0.3	1.00

Top 6 highest correlated variables were selected to represent the correlation matrix. Other variables had a score lower than 0.20, and were deemed not significant. The only negative correlations were observed between Box Office gross and R content rating -0.25, next to the Comedy genre -0.11.

As expected the production cost was moderately correlated to Box Office gross with the coefficient 0.63. This indicates a positive relationship between gross and budget: the higher the budget, the higher the gross of a movie. Movies with a higher budget have a large advertising budget, and can reach more audiences. They also feature popular actors and actresses, and costly special effects. Number of users voted was also moderately correlated with the Box Office gross with the coefficient 0.62, followed by the number of user reviews (coefficient 0.54), and the number of critic reviews (coefficient 0.45). Popularity of a movie on the Facebook social network, as well as the movie score had a weaker correlation to Box Office gross.

Budget

As stated above, the budget shows the highest correlation to the Box Office gross, which was expected. The significance value lies in the moderate level. The average mean value is $38.34M, while the median value is $25M. The movie with the highest budget is ‘Pirates of the Caribbean: At World's End’ with a $300M budget, but with a very low return of 3.13$. The movie with the lowest budget is ‘Tarnation’ with a $218 budget, and $0.59M gross. Its ROI is expectedly very large: 271,466%. The assumption is made that the dataset contains a certain number of anomalies such as low budget movies with high profitability, which influences the relationship between the budget and gross.

The dataset showed that among the 12 highest budget movies had production costs of $230M and higher. The highest grossing movie in this sample was again ‘Avatar’. Its production cost lies at $237M. The most dominant genres among the top 12 highest budget movies are Action and Adventure.

ROI

In terms of profitability, ROI seems to show how spread out the results are. The average mean value of ROI is 540.4%, while the median lies in the negative with -49.1%. This obvious difference indicates a certain number of extreme cases or anomalies, as well as low profitability of movies in general.

> Table 3: Top 5 movies with the highest ROI <

No.	Movie Title	ROI %	Gross ($) Millions	Budget ($)	Genre
1.	Paranormal Activity	719348.5	107.92	15000.0	Horror
2.	Tarnation	271466.0	0.59	218.0	Biography
3.	The Blair Witch Project	234116.8	140.53	60000.0	Horror
4.	The Brothers McMullen	40886.4	10.25	25000.0	Comedy
5.	The Gallows	22657.8	22.76	100000.0	Horror

Among the top 12 movies with the highest ROI, a couple of things stand out. Firstly, the mean budget is $0.37M, and the highest value is $2M, while the average gross is $64.0.M, with the highest value of $184.93M. Secondly, the average ROI value among the top 12 is 114542%, which in terms of profitability is pretty significant. Lastly, the common genres present in the top 12 movies with the highest ROI are Comedy movies, followed by the Horror and Drama genre.

Number of users voted, number of reviews, number of critic reviews

Results showed a significant correlation of number of users voted and the Box Office gross, while the number of reviews, as well as the number of critic reviews seems to be not as significant. The average mean value of the number of users voted is around $100K, while its median value is around $500K. Mean values of the number of critic reviews and the number of reviews is 162 and 330 respectively. The highest number of votes is $1.68M for 'The Shawshank Redemption', with an average production cost of $25M, and gross $28.3M. The second highest number of votes is $1.67M for 'The Dark Knight', with a much higher production cost of $185M, and gross $533.32M. Mean values for the top 12 movies with the highest number of votes are $1.3M, and median value $1.24M. Average gross is $259.5M, production cost $91.3M, and average ROI 432.5%. Number of reviews, as well as the number of critic reviews average at 3118, and 368. The dominant genre is again the Action genre.

Social Media Popularity

Social Media popularity measured in likes on Facebook showed some significance in correlation to the Box Office gross, however the significance was lower, and it differed among the popularity actors have, the cast, and the movie itself. The average number of likes for a movie is 9053, for the cast 11655, and for the lead actor 7873. The average budget for the top 12 movies with the highest number of likes is $140M, while the budget is $50M average when we look at the top 12 movies sorted by the highest number of likes for the lead actor. When looking at the statistics for the top 12 movies sorted by the highest number of likes for the cast, we see the average budget is $88M.

Modeling

An assumption was made that the data showed a fairly linear relationship, therefore the model used for prediction was Multiple Linear Regression. A number of features were used to build the model, based on the correlation matrix previously displayed. The Model showed significance levels for the following features: budget, number of users voted, Facebook likes of the cast, Facebook likes for both lead actor and the supporting actor, as well as duration and title year. Details of the model's results can be seen in the table below.

> Table 4: Determinants of Box Office gross <

	Coefficients	t-value	P-value
Intercept	4.012e-17	3.64e-15	1.000
Budget	0.47	35.06	0.000
Number of users voted	0.41	19.38	0.000
Facebook Likes Cast	0.55	6.48	0.000
Actor 1 Facebook Likes	-0.48	-6.67	0.000
Actor 2 Facebook Likes	-0.14	-5.09	0.000
Duration	-0.06	-4.82	0.000
Title Year	-0.09	-7.28	0.000
R squared	0.586		
F-value	437.9		

The model showed a significant relationship between the budget and gross, the strongest among other features, which was expected. Another major contributor to the movie's gross is the number of users who voted.

Conclusion

The most significant contributor to the Box Office gross is the budget. This result is consistent with other studies. However, the budget itself cannot be used as a certainty for predicting a movie's Box office gross, as there are a large number of extreme cases in terms of profitability. The results showed a different picture when the focus was on the profitability of a movie, rather than its gross. This indicated a large unpredictability of the audience's reaction, which is consistent with other studies. Larger budgets reach larger audiences, and feature cast which is popular on Social media networks. Linear relationship between the gross and the number of users voted indicates the audience's reaction to a movie, but in terms of movie’s production it cannot be used as a determinant as it is a reaction that follows the release. The most dominant genre was the Action genre, however the model showed no significant relationship between the movie’s genre and its profitability.

Some recommendations are in place for future consideration: the results showed a high number of sequels present in the top 12 grossing movies, it might be significant to explore the relevance of this relationship, and how it affects the Box Office gross. It is safe to assume that audiences

would be fairly familiar with the franchise, and would make an informed decision about seeing the sequel. It is worth noting that a sequel would only be considered if the previous movie was a hit.

It is also worth noting that other studies showed a significant relationship between the number of nominations a movie received and its gross, as well as if the movie was released by a major studio.