A More Scientific Approach to Profitable Movie Making

Project Proposal

Prepared for: Metis Staff Instructors: Chris Bruehl, Kevin Birnbaum
Prepared by: Rahul D. Raju
July 28, 2021

EXECUTIVE SUMMARY

Objective

In the 21st century, institutions and industries across the globe have embraced the use of big-data in their decision making processes. Each of these organizations and industries face unique challenges which in turn affects the use and efficacy of this data. One industry with particularly interesting challenges is film making. Hollywood faces the dual task of not only making "good" movies but ensuring they are profitable as well. This can be somewhat difficult as the two factors are not always correlated. Also, the notion as to what makes a "good movie" in the pre-production phase can be highly subjective. As most if not all major studios now belong to publicly listed media conglomerates, a significant emphasis is now being placed on the profitably component of this equation. Particularly with the stock market at all time highs, movie industry stocks could be susceptible to anything that diminishes the bottom line. Thus studios must explore every possible avenue in ensuring that the spectacular losses of the past are not repeated or at the very least minimized.

The table below represents the biggest box office losses of all time, ranked in descending order.

Biggest box office bombs

Title	Year	Net production budget (millions)	Worldwide gross (millions)	Estimated loss (millions)		Ref.
				Nominal	Adjusted for inflation	[nb 1]
John Carter	2012	$263.7	$284.1	$114–200	$129–225	[# 60]
The Lone Ranger	2013	$225–250	$260.5	$160–190	$178–211	[# 67]
Mortal Engines	2018	$110	$83.7	$174.8	$180	[# 73]
King Arthur: Legend of the Sword	2017	$175	$148.7	$114–153.2	$120–162	[# 63]
Battleship	2012	$209–220	$303	$150	$169	[# 14]
Tomorrowland	2015	$180–190	$209	$90–150	$98–164	[# 103]
Pan	2015	$150	$128.4	$85–150	$93–164	[# 79]
Mulan	2020	$200	$66.8	$147	$147	[# 74]
Mars Needs Moms	2011	$150	$39	$100–144	$115–166	[# 69]
Onward	2020	$175–200	$142	$135	$135	[# 77]
Dark Phoenix	2019	$200	$252.4	$79–133	$81–137	[# 27]
A Wrinkle in Time	2018	$125	$133.4	$130.6	$135	[# 111]
Terminator: Dark Fate	2019	$185–196	$261.1	$110–130	$113–134	[# 101]
The 13th Warrior	1999	$100–160	$61.7	$69–129	$107–200	[# 1]
Sinbad: Legend of the Seven Seas	2003	$60	$80.8	$125	$176	[# 92]

Solution

With so much money at risk, a much more scientific approach is needed with respect to idea generation, movie selection and development. One approach, could be the use of an algorithm such as *Linear Regression*. The algorithm would employ the plethora of data available and use statistical methodologies to identify patterns within that data in order to build a predictive model. The model would aid us in identifying characteristics that correlate with movie profitability, specifically world wide box office gross. Movie pre-production variables could be adjusted to reflect the characteristics(features) of past profitable movies. Various statistical metrics will also provide the ability to evaluate the efficacy of such a model. Finally, assuming it's efficacy, this model could also be used in providing a quantitative component in pitches & presentations to potential institutional film investors.

Project Data

The primary source of data will be attained by web-scraping information from boxofficemojo.com. Each row of data will contain financial and production features of each movie within the data frame. At least 1000 of the most profitable movies will be sourced. Supplemental data will be sourced from IMBD.com as well.

Project Tools

1). Web Scraping and Parsing tools: BeautifulSoup, Selenium

2). Data Cleaning & Analysis: Pandas

3). Data Visualization: MatPlotLib, Seaborn

4). Statistical Analysis: Sci-Kit Learn, StatsModels

Minimum Viable Product Vision

A MVP for this project would consist of the following:

1. A list of all the feature(independent) variables to be used in this analysis.
2. A basic pair plot showing, amongst other things, if a linear relationship exist between feature variables and the target variable or amongst the features themselves.
3. A correlation table/heat map showing the strength of any linear relationships between feature variables and the target variable or amongst the features themselves.
4. A baseline regression model of key features with associated statistical output.