What makes a Hollywood film successful?

Joanna Andari, Karim Awad, Jiye Ren, Nirbhay Sharma

8 October 2017

```
## Warning: package 'bindrcpp' was built under R version 3.4.2

## Warning: Too few values at 545 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
## 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...

## Warning: Column `director_name`/`Var1` joining character vector and factor,
## coercing into character vector

## Warning: Column `actor_1_name`/`Var1` joining character vector and factor,
## coercing into character vector

## Warning: Column `actor_2_name`/`Var1` joining character vector and factor,
## coercing into character vector

## Warning: Column `actor_3_name`/`Var1` joining character vector and factor,
## coercing into character vector

## Warning: Column `director_name`/`Var1` joining character vector and factor,
## coercing into character vector
```

Section 1: Introduction

The film industry is expected to generate global box office revenues of almost $49.3bn by 2020, from a forecast of US$38.3bn in 2016[1], representing a CAGR of over 32%. These substantial growth levels are likely to result in vast resources being poured into the industry, in search of the next "big hit". Countries, recognising the value of developing a film industry to support a variety of local and national initiatives, have supported these activities, providing tax breaks to investors, to mitigate the material upfront costs involved.

Outside of India, the US continues to lead the way in producing a substantial volume of films, accounting for an estimated US$10.7bn in box office revenues in 2016[2]. Yet in contrast to Indian films, which principally serve a domestic audience, Hollywood films benefit from a global distribution network. This results in films having to cater for different audiences in different countries, overcoming issues varying from political censorship through to cultural sensitivities, to maximise their appeal and potential commercial success.

With global audiences in mind, are there factors that can help predict a Hollywood film's popularity, and with that, guide future decisions on the types of films produced? In turn, can this better inform investors on potential risks involved with backing particular projects?

This report seeks to examine a random sample of movie data pulled from IMDB data between 2011-16, concentrating on the US given its role in influencing the wider film industry. This will seek to understand how we should measure popularity, focusing on 2011-15 data, before devising a regression model that can help predict popularity. We will then test its capacity to predict 2016 trends, before noting the limitations of our work, and areas for further development.

As this report shall conclude, [...]

[1] Statista (2016)
[2] Statista (2016)

Section 2: Data sampling and cleaning

The data compiled for this project/analysis is obtained from 2 different datasets:
* **IMDB Movie Dataset**
* **Oscar Awards Dataset**

We have performed data cleaning on both of these datasets and then compiled them to a single dataset in order to obtain the 'working data'. Below is a summary for the data cleaning operations performed on both of these datasets.

1. **IMDB Movie Dataset:** The raw data consisted of 5044 observations, containing details of movies from the IMDb database over 28 variables. These variables included information about the movie performance on the box office, the actor(s) name, direcor(s) names, movie genre, movie facebook likes and more.
    - **Step 1 - Cleaning up inconsistent observations:** The raw dataset contained a lot of inconsistencies, namely 'NA' or Blank values for several observations. In order to get a consistent data to perform analysis on, we ***removed all the observations with any NA or blank values***.
    - **Step 2 - Removing duplicate values:** We observed that the dataset contained duplicate observations, which were also removed.
    - **Step 3 - Limiting our dataset:** In order to do a comprehensive analysis, having a dataset of movies belonging to a single country or industry will be a good place to start. Furthermore, ever since the the outset of facebook and other social media, the success of a movie is heavily reliant on its social presence and digital marketing. Therefore, to completely account for these factors, and also ensure that the data we analysed is consistent, we have limited our data to ***movies made in the USA after the year 2011***.
    - **Step 4 - Segregating the movie genres:** The dataset contained the movie genres in the form of a single string, with every genre of the movie, delimited by a '|' symbol. In order to get the actual genres separated and ready to use, we used created flags in order to check whether the movie is of a particular genre. For example: the 'action_genre' column will have a 1 value for an observation, if it is an action movie.
    - **Step 5 - Calculating the number of occurences of each actor/director:** Another interesting variable that we expect to be of significance to our analysis was the number of times an actor/director has showed up in our dataset. That data would allow us to possibly find interesting points on how an actor/director can impact a movie's performance/success. Thus, we added variables/columns with the ***number of occurences for actors and directors***.
    - **Step 6 - Calculating the movie ROI:** We add another variable, 'movie_roi' which contains the ***return on investment for each movie***. This will help us in our analysis.
    - **Step 7 - Selecting only the required columns:** The raw dataset consisted of 28 variables, many of which are not required in our analyses. Thus, we remove the unwanted columns and keep only the data that is specific to our analysis. In the process we reduce the size of the dataset leading to better performance. (See data description for more detials).
      The cleaned movie dataset now has 535 obseravations over 39 variables. Please note that there are 21 variables which are genre indicators, and 4 are occuence variables.
2. **Oscar winners dataset:** The raw dataset of awards contained 2321 observations over 4 variables. This database consisted the nominees and winners of the Academy Awards (or Oscars) ever since its inception, i.e. from 1927 to 2015.
    - **Step 1 - Cleaning up inconsistent observations:** The raw dataset some inconsistent observations where the name of the winner and the film was placed in the incorrect column. Since this was the case for the initial few records only, we ignored this and copied the values in the 'film' column to the 'names' column.
    - **Step 2 - Limiting our dataset:** The dataset included both nominess and winner and from every category. We were interested in collecting data about the actors, directors, and the best picture

winners only in order to see how that impacts the success of a movie. Therefore, we ***excluded the nomiees and only kept the winners for the awards*** related to Best Actor/Actress in a Lead/Supporting role, Best Director and Best Picutre.

- **Step 3 - Counting the awards:** In order to get the data of the number of wins for each individual, we introduced a new column 'num_wins' which counted the ***number of occurences of a name in the winners list***. Using this data makes sense as in order to asses the impact of this on a movies popularity/success.
- **Step 4 - Selecting the required columns:** The raw dataset contains 4 variables, and a count column, out of which only the 'count' and 'name' column is required for our analysis. Thus we select only a small subset of the raw data for our analysis.
  The cleaned awards data now has 434 observations over 2 variables. This data will come handy to check which movies/actors/directors have oscar(s) to their name(s) and how that impacts the success of a movie.

3. **Final (Cumulative) Dataset:** In order to get the final working database, we combine the information from the two cleaned datasets.
   - **Step 1 - Summarizing the datasets:** We need to get the number of award wins for actors, directors, and movie for each observation in the into the IMDB movie dataset, from the 'Oscar winners dataset'. We do this by introducing new variables that captures ***the count of oscars for each actor, director, and movie for each observation***.
   - **Step 2 - Diving the final data into two datasets:** The entire basis of our analysis was to create a model that can predict the success of a hollywood movie. To achieve that, we will try to build a model (or hypothesis) using the data for movies from 2011 to 2015 and then test the model (or hypothesis) for the movies in 2016. This will give us a clear indication of how our model fits with current and future data. Thus we ***create two subsets of the final data, one for data of movies from 2011 to 2015 and the other with the data of movies for 2016***.

Section 3: How do we measure a film's popularity?

What metrics should we use to determine our dependent variable?

Determining what variable should represent popularity, itself a subjective quality, gives rise to several potential proxy variables. These can be divided between those that are financially driven relative to social media metrics, of which we consider each in turn, arriving at five possible variables.

An obvious proxy would be revenues, should we determine this a valid measure of commercial success. Contrary to the thoughts of those "voting with their wallets", this is not entirely a straightforward metric. A limitation of the data available is we are only able to track box-office ticket sales, rather than all associated revenues. This ignores digital sales from streaming, DVD sales, and any other merchandise or commercial agreements (e.g. in-film advertising).

An example of this shortcoming would be the motion picture, "The Interview". Although extreme given the political sensitivities with North Korea, which led to a limited global release, "The Interview grossed $40m in digital rentals... and earned an additional $11.2m worldwide at the box office on a $44m budget[3]". By relying on box office revenues alone, we may not get a complete view on how commercially successful a film has been. This is something we discuss further as a limitation within our work, but which we do not explore further.

A second issue is whether a low grossing film is still a success, relative to its underlying budget? This may be more applicable to arthouse films, rather than a typical Hollywood blockbuster, but is something we are able to condition against by examining a film's return on investment; defined by its box office revenues divided by its total budget.

Social media provides us with three relevant metrics. The first surrounds the number of facebook likes a film has received. Although a singular expression of popularity, this provides some indication of fan reaction, premised on any "likes" resulting from watching the film, as opposed to being based on trailers or other

[3]Wikipedia (2017)

promotional material. We are mindful this may not be entirely accurate, as we do not know at which point this data was collected (e.g. how many days / weeks after the film was released in the US, and how the latter is likely to vary by geography).

The remaining two metrics concern IMDB scores, and the number of votes received. IMDB scores should provide a direct measure of popularity, with our initial assumption based on this most likely to be normally distributed (i.e. this theoretically should capture a broad range of views between a ranking of 0-10). The number of users may also help quantify the depth of support, and depending on the distribution of IMDB scores, represent a useful proxy for capturing popularity.

Descriptive analysis

We produce boxplots and density graphs of the five variables above, along with budget, to provide a sense-check. These can all separately be found in the appendices.

We summarises some of the key descriptive findings of these graphs, along with illustrating the results of our ROI boxplot and IMDB density graph:

```
stargazer(data11_15, type="latex", title="Table with Stargazer")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Oct 08, 2017 - 20:54:23

```
# imdbden <- ggplot(data=data11_15)+geom_density(aes(x=(imdb_score)))
# roibox <- ggplot(data=data11_15)+geom_boxplot(aes(x=1, y=return)) + ylim(0,5)
#
# grid.arrange(imdbden, roibox,  ncol=2)
```

Correlation analysis

```
## Warning: package 'knitr' was built under R version 3.4.2
```

	imdb_score	movie_facebook_likes	gross	return	budget	num_voted_users
imdb_score	1.00	0.47	0.29	-0.07	0.15	0.53
movie_facebook_likes	0.47	1.00	0.53	0.01	0.38	0.80
gross	0.29	0.53	1.00	0.04	0.68	0.66
return	-0.07	0.01	0.04	1.00	-0.25	-0.03
budget	0.15	0.38	0.68	-0.25	1.00	0.52
num_voted_users	0.53	0.80	0.66	-0.03	0.52	1.00

The most surprising result is the lack of correlation with the ROI variable.

[skewness]

Table 1: Table with Stargazer

Statistic	N	Mean	St. Dev.	Min	Max
title_year	498	2,012.948	1.382	2,011	2,015
duration	498	112.823	19.861	72	240
gross	498	80,765,853.000	89,529,230.000	1,332	652,177,271
num_voted_users	498	156,491.100	154,743.000	22	1,144,337
num_user_for_reviews	498	367.480	349.370	1	2,725
budget	498	59,531,865.000	58,411,495.000	9,000	263,700,000
imdb_score	498	6.592	0.887	1.600	8.600
movie_facebook_likes	498	38,192.020	35,800.310	44	349,000
return	498	2.341	3.950	0.000	53.250
actiondummy	498	0.313	0.464	0	1
adventuredummy	498	0.275	0.447	0	1
animationdummy	498	0.074	0.263	0	1
biographydummy	498	0.080	0.272	0	1
comedydummy	498	0.347	0.477	0	1
crimedummy	498	0.179	0.383	0	1
documentarydummy	498	0.020	0.140	0	1
dramadummy	498	0.442	0.497	0	1
familydummy	498	0.124	0.330	0	1
fantasydummy	498	0.183	0.387	0	1
historydummy	498	0.026	0.160	0	1
horrordummy	498	0.104	0.306	0	1
musicaldummy	498	0.026	0.160	0	1
mysterydummy	498	0.102	0.303	0	1
newsdummy	498	0.000	0.000	0	0
romancedummy	498	0.147	0.354	0	1
scifidummy	498	0.181	0.385	0	1
sportdummy	498	0.038	0.192	0	1
thrillerdummy	498	0.293	0.456	0	1
wardummy	498	0.024	0.154	0	1
westerndummy	498	0.012	0.109	0	1
director_occurence	498	1.988	0.245	1	4
actor1_occurence	498	4.086	0.936	1	12
actor2_occurence	498	3.916	0.565	1	6
actor3_occurence	498	7.616	1.520	1	10
dir_oscars	498	0.056	0.231	0	1
movie_oscars	498	0.056	0.231	0	1
actor_oscars	498	0.211	0.481	0	2

Appendix 1 - Glossary of data definitions

Appendix 2 - Boxplots and Density graphs

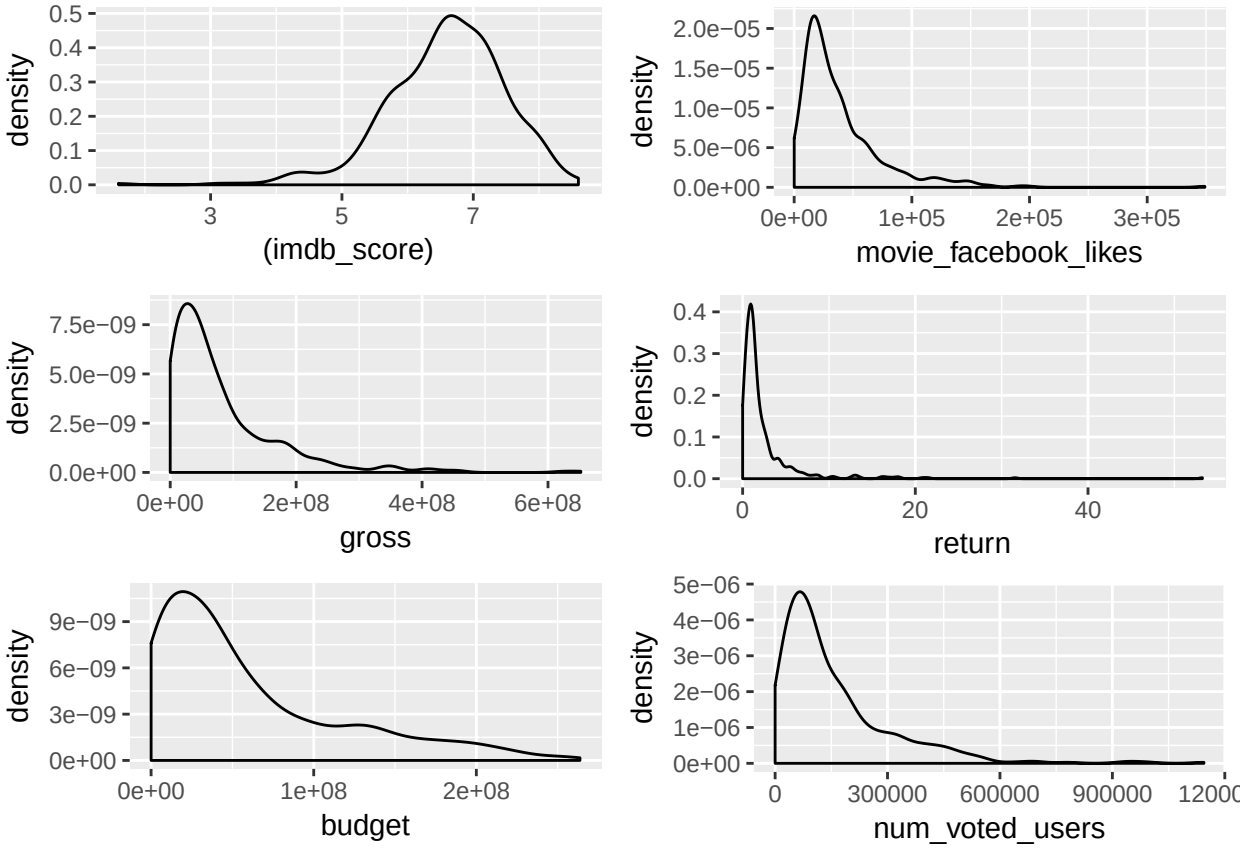
```
## Installing package into 'C:/Users/karim/OneDrive/Documents/Imperial - Business Analytics/Maths and S
## (as 'lib' is unspecified)
```

```
## package 'gridExtra' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\karim\AppData\Local\Temp\Rtmp4OP3yy\downloaded_packages
```

```
## Warning: package 'gridExtra' was built under R version 3.4.2
```

```
##
## Attaching package: 'gridExtra'
```

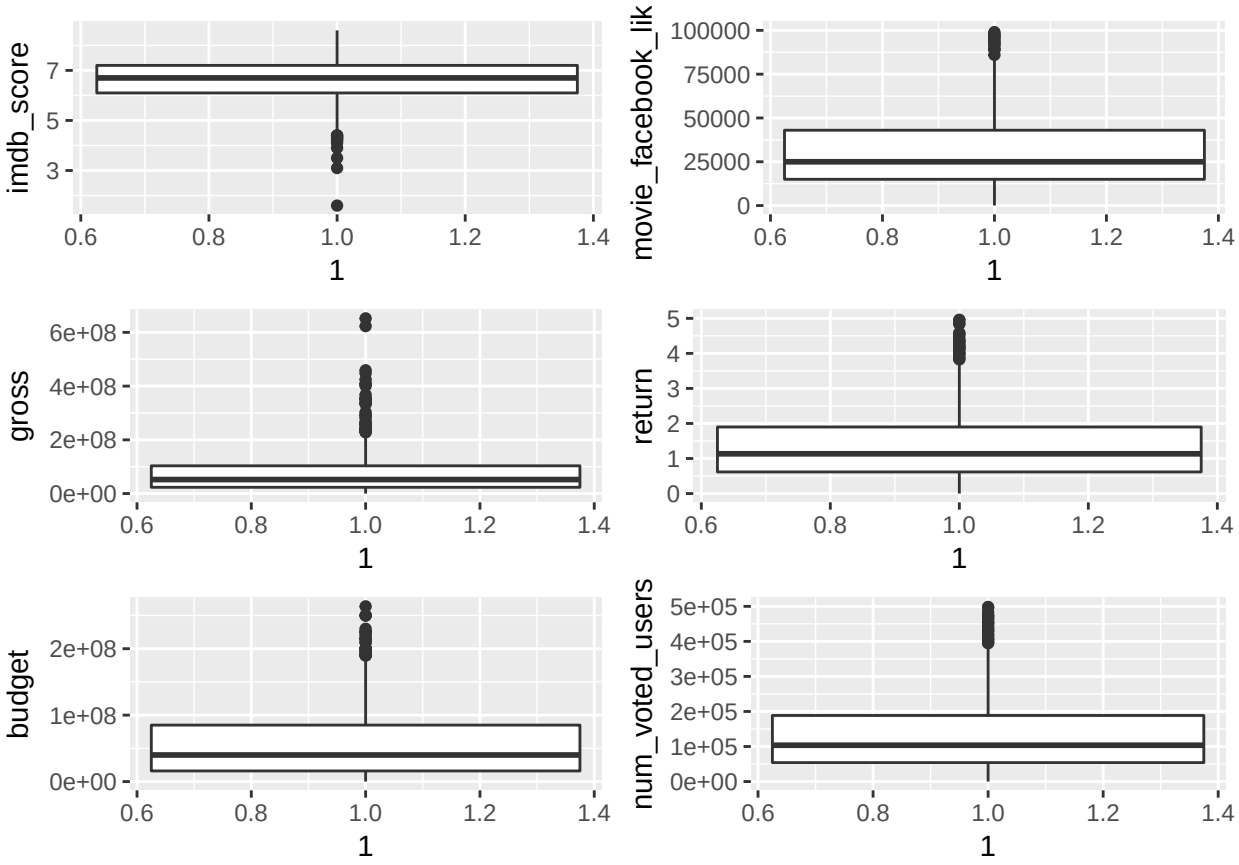
```
## The following object is masked from 'package:dplyr':
##
##     combine
```




```
## Warning: Removed 30 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 50 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 16 rows containing non-finite values (stat_boxplot).
```


imdb_score
7
5
3
0.6
0.8
1.0
1.2
1.4
1
movie_facebook_lik
100000
75000
50000
25000
0
gross
6e+08
4e+08
2e+08
0e+00
return
5
4
3
2
1
0
budget
num_voted_users
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00

Bibliography and references

We have made reference to various forums on coding best-practices, notably including Stackoverflow.com, and other relevant sources of R studio information available online. These have been cited during the main document, where they have been specific to our study, or represent the application of coding practices not within the Maths and Statistics Foundations for Analytics module. We have not referenced general R studio packages applied within this module (e.g. dplyr; stargazer; etc).

We obtained the original dataset, in conjunction with information on Oscar academy awards, from the following websites: https://www.kaggle.com/tmdb/tmdb-movie-metadata/data https://www.kaggle.com/theacademy/academy-awards

In terms of documents we have referenced:

[1-2] "Film industry in the US", Statista, Dossier (2016) https://www.statista.com/study/11472/film-industry-in-the-united-states-statista-dossier/ [3] "The Interview", Wikipedia page (2017) https://en.wikipedia.org/wiki/The_Interview#Pre-release_reaction