



Movie Box Office Revenue Prediction

ABSTRACT

Predicting the gross of the movie based on various factors known at the start of filming

GROUP 3 - SHERLOCK

Adithya Selvaganapathy
A0186084X

Dinesh Kumar Agarwal Vijayakumar
A0186283W

Hemanand Moorthy
A0186104L

Mookkandi Sathan Karthikeyan
A0186448N

Spatika Narayanan
A0088416X

Swetha Narayanan
A0074604J

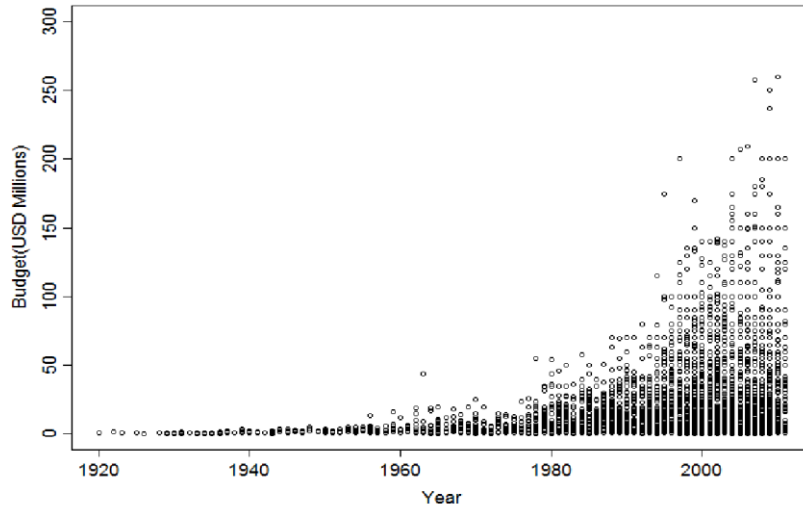
BT5153 :

TOPICS IN BUSINESS
ANALYTICS

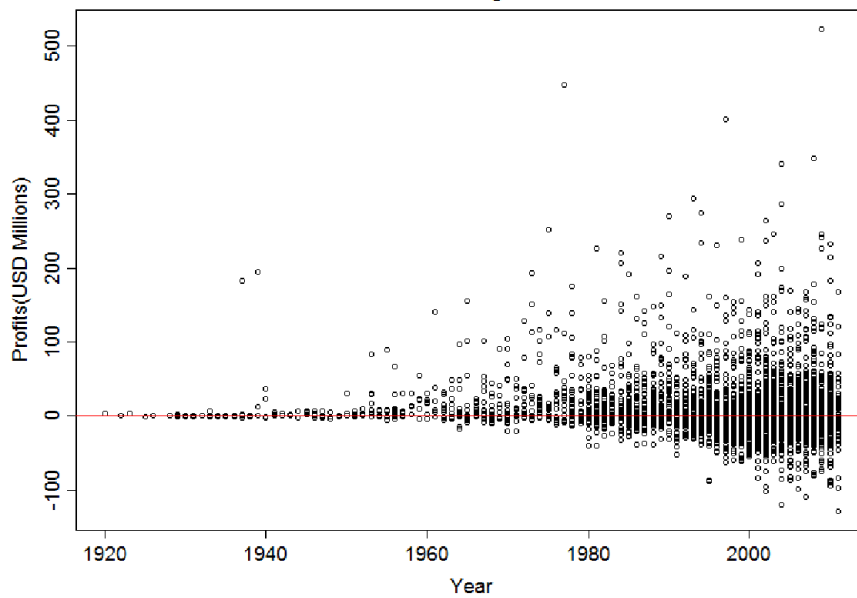
Problem Statement

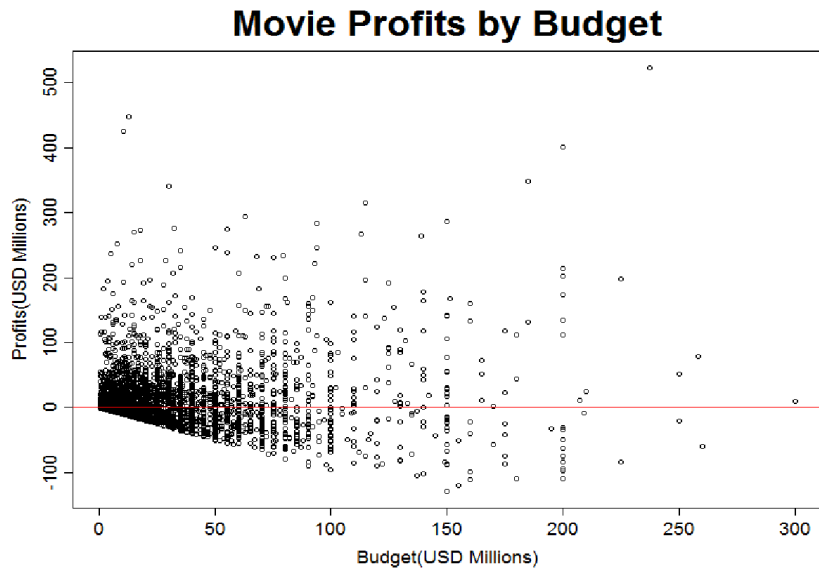
The global box office revenue is projected to scale to 50bn USD by 2020. The fact that movie making is an expensive business, makes it extremely imperative to ensure that the movies experience box office success.

Budgets by Year



Profits by Year





As observed from the graphs above, even though production costs are on an increase, the same cannot be said about the ROI. This makes movie production a risky business to be involved in.


Our aim is to build a prediction model which considers factors which are known before the movie making starts to predict its worldwide box office revenue. To add on to this, we would also be displaying a list of movies having a storyline similar to the movie which is to be produced.


Motivation

The success of the movie depends on a large set of factors, for example:



1. The popularity of the cast and crew
2. The genre of the movie
3. The effectiveness of the promotion events
4. Initial release reviews (reviews from premiere shows till first weekend after release) and post release reviews (reviews after two weeks of release)

The data about the movie is often not found in a single source. For example, consider a movie page from IMDb:

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE  SHARE

+ Toy Story (1995) ★ **8.3**
772,883  Rate This

G | 1h 21min | Animation, Adventure, Comedy | 22 November 1995 (USA)






1:02 | Trailer 11 VIDEOS | 128 IMAGES

A cowboy doll is profoundly threatened and jealous when a new spaceman figure supplants him as top toy in a boy's room.

Director: John Lasseter
Writers: John Lasseter (original story by), Pete Docter (original story by) | [6 more credits »](#)
Stars: Tom Hanks, Tim Allen, Don Rickles | [See full cast & crew »](#)

[+ Add to Watchlist](#)

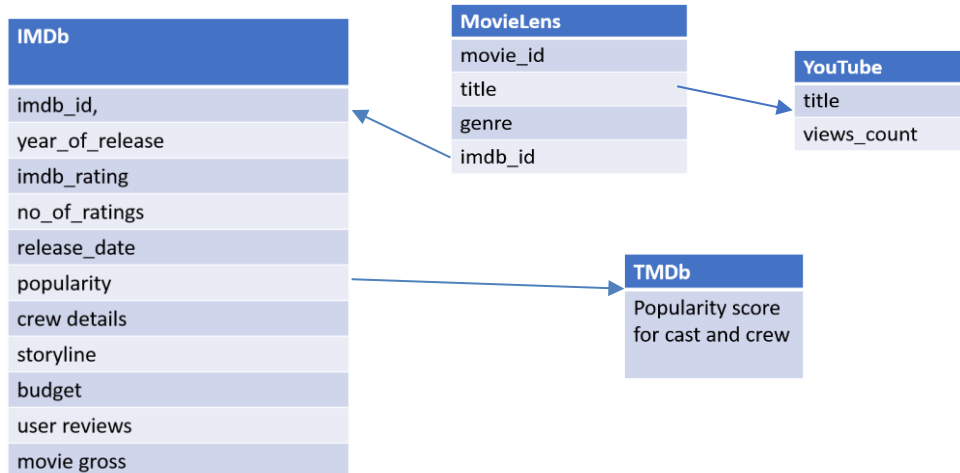
 Metascore | [Reviews](#) |  Popularity

This gives us information about the cast/crew and the genre the movie belongs to. However, it fails to talk about the popularity of the cast/crew, how the reviews influence a box office collection and the effectiveness of promotion events.

Working on predicting the box office success or failure is not new. For example, there are Kaggle competitions which have attempted to solve the same problem (<https://www.kaggle.com/tmdb/tmdb-movie-metadata/data>).

However, our aim is to not limit our model to a single source but to scrape information regarding the movie from diverse sources to get a holistic view of the various factors influencing box office revenue.

Prospective Data Sources



DATA SETS	SOURCE	DATA COLLECTION METHODS	DESCRIPTION
Movie_data	<i>IMDb</i>	Web scraping the IMDb web pages to extract information on the movie and its cast, using Python's 'beautifulsoup' package	Information on a movie including the user review extracted from IMDb
Cast_data	<i>IMDb</i>		Information on the cast, financials, storyline and box office extracted from IMDb
Movies	<i>MovieLens</i>	Data collected by a research group 'GroupLens' from the University of Minnesota. The MovieLens dataset was collected over time from an online survey engine hosted by GroupLens that surveys MovieLens users. The dataset was obtained for MovieLens' recommendation system	A dataset of movie_id, title and the movie's genre
Links	<i>MovieLens</i>		URLs to the IMDb and TMDb pages for each movie
Movie popularity	<i>YouTube</i>	Using Google API calls to pull the number of views for the movie trailers	The number of view for the trailer will act as a proxy for the success of promotional activities
The popularity of cast and crew	<i>TMDb</i>	Using TMDb API calls to pull the popularity score for the cast and crew	TMDb has arrived at a score using the number of likes on their Facebook pages, Twitter following, age, and recent box office performance. This score can be used for determining the current popularity of the cast and crew

Possible Text Mining Methods

1. Using NLP on two categories of user reviews extracted from IMDb. We would be taking user reviews made till the first weekend of movie release and reviews made after the first weekend of release to arrive at two different sentiment scores.
2. Using NLP on the storyline of the movies extracted from IMDb. Cosine similarity to arrive at ten movies having a similar storyline to the movie in question

How the Model can be Used

1. Predicts the box office performance with a set of known input factors like actors, cast, crew, and genre
2. User reviews and the success of the promotional events cannot be predicted before the movie is made. So, the two features from the user reviews mentioned above can be simulated by the producer to know how revenue varies. For example, the producer could fix the scores to their minimum possible values (highly negative user reviews), and repeat for maximum possible scores (highly positive user reviews), to generate the range of possible box office revenues. Similarly, the number of trailer views from YouTube trailers/teasers can be simulated.
3. Producers get to know ten movies with similar storylines. This would enable them to learn from the good/bad from those movies, their performance and could plan their budget accordingly.

Shortcomings

Predicting the success/ failure of a movie is a qualitative problem rather than a quantitative problem. Even if the past set of directors, cast, and crew had given a great box office movie it is not necessary that the next movie with the same team would lead to similar performance. Some factors like time of release, people's mindset during that time (example, calamities occurring during that time can hamper collection) and competition from other movies are some factors which we cannot address using our model.