Predicting Box Office Success for Movies

Ashish Mohapatra
axm160031@utdallas.edu

Dhruv Gajanan Sangvikar
dgs160230@utdallas.edu

Vipul Nataraj Jarmale
vxj160130@utdallas.edu

***Abstract*—The aim of this project is to predict the box office success of an upcoming movie using the data available about the movie online and previous box office data. How the movie performs at the box office is a combination of various factors like the actors in the movie, the directors, and producers of the movie, the views it is getting on YouTube, the film rating given by MPAA and many other factors. We aim to predict using a combination of some or all of these factors using linear regression.**

***Index Terms*—Movies, Movie-Revenue, Youtube, OMDB, Scrapper, IMDB, Revenue**

I. Introduction

The internet is a great source of movie information due to which a lot of people visit various sites like IMDB, YouTube, Twitter, Facebook to check-out information about the movies, the actors and directors of those movies and a sneak peek about what the upcoming movie is about. An enormous volume of information about what people think about the movie is available online which can be gathered and can be put to good use. One such example is the number of views it is raking up on YouTube. The more views a movie gets online as the movie nears its release date, greater the buzz it is generating which can greatly contribute to the box collections of the movie. Another factor that contributes is the combination of actors in the movie. An award-winning director or an actor who is known for great performances on screen, a good IMDB rating will definitely kindle interest in the movie-goers. And sometimes, it's not just one actor but a combination of actors can be crucial to movie's success. Movies like Avengers, Justice League are classic examples. The rating given by MPAA also plays a role in the number of people who will watch the movie which can influence the collections of the movie. Though this can be a little less obvious than the above-mentioned factors, analyzing our data can give us a clearer picture as how to much it would influence by itself alone and how much say it would get when considered in combination of others factors. This type of prediction will help producers and studios make financial decisions as to which actors or a combination of actors to hire for a movie and which directors' movie ideas to fund, the MPAA rating the movie they should be looking to get etc.

II. Dataset Aquisition

To begin, we aggregated our dataset primarily from the Internet Movie Database (IMDB) and the Open Movie Database (OMDB). We acquired a text representation of movie title from IMDBs database (limiting our search to movies from 2000 and later due to a lack of information available for earlier movies)

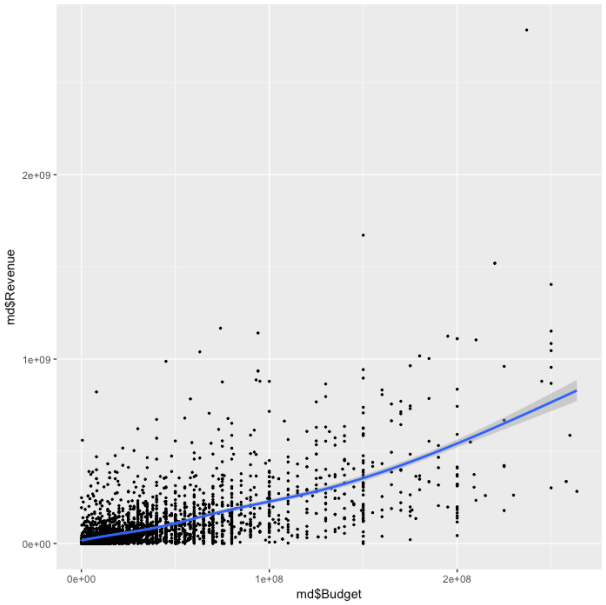

Fig. 1: Budget vs Movie Revenue

A. *Extraction of data from OMDB*

In order to get metadata and box office revenues, We hit the OMDB API with the help of JSON library for each of the films and aggregated the metadata for each film into a single database (MovieDeatil.csv). The entire process took roughly ten hours to complete. Once the metadata was extracted for each film, We filtered the list of films to films that had valid box-office data, removing all the NULL values and noises in the process. Of all films in our list, only 3,500 of them had valid data.

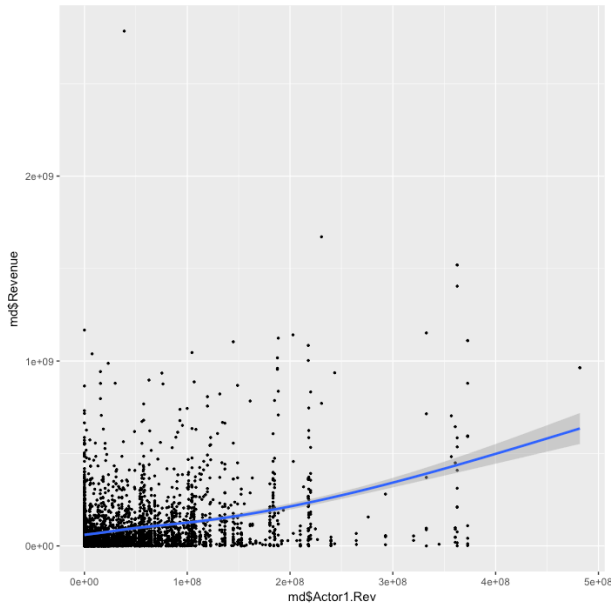

Fig. 2: Lead Actor Revenue vs Movie Revenue

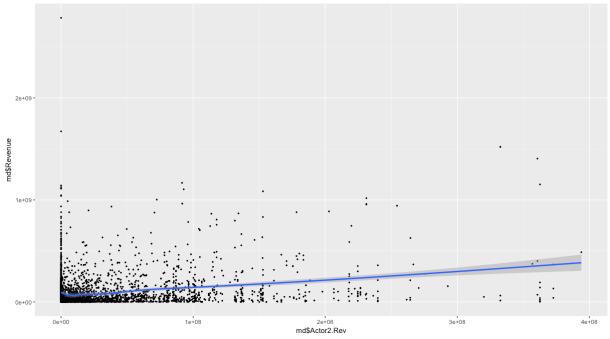

Fig. 3: Lead Actor 2 Revenue vs Movie Revenue

B. *Extraction of data from YouTube*

We also extracted Likes, Dislikes and View count of the movies trailers from YouTube, in which we have used YouTube Data API v3. We searched for the keyword Official Trailer along with movie year. We picked top three results from the Output and took summation of all three results for likes, dislikes and view count. We aggregated the metadata into another dataframe (YoutubeDetails.csv). We filtered out the invalid data.

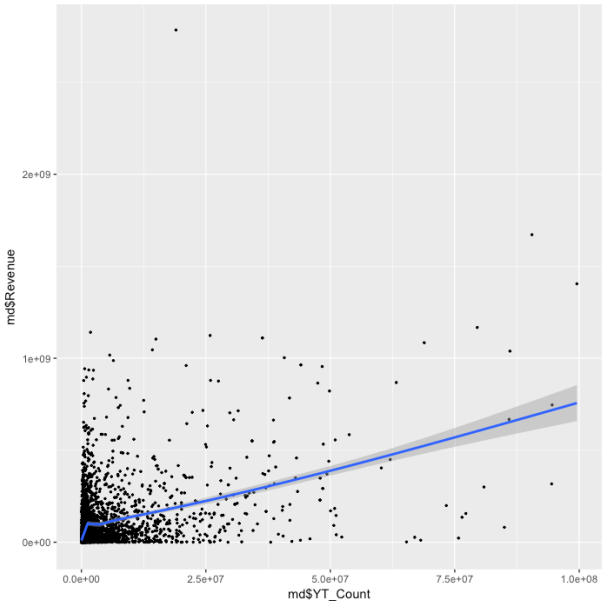

Fig. 4: Youtube view Count vs Movie Revenue

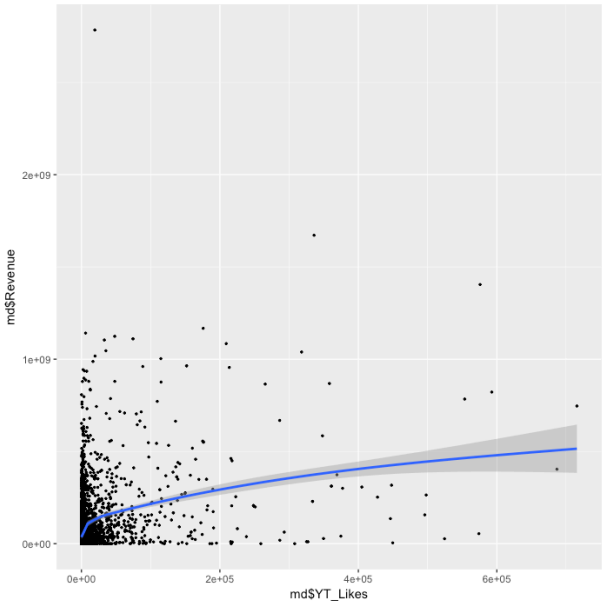

Fig. 5: YouTube Likes vs Movie Revenue

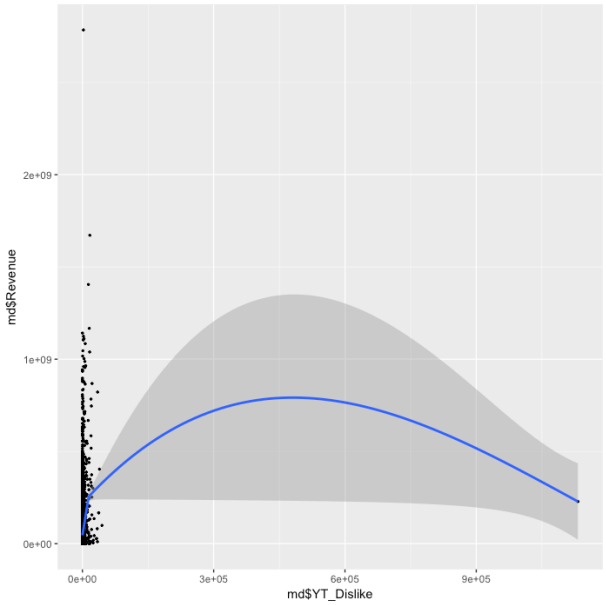

Fig. 6: YouTube Dislikes vs Movie Revenue

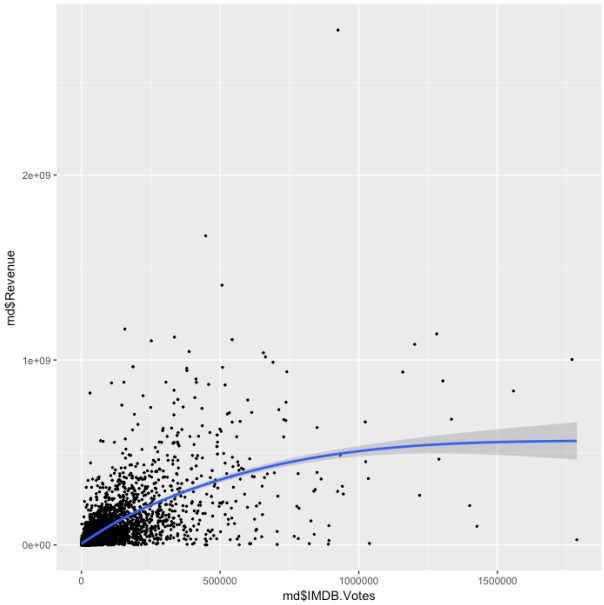

Fig. 7: IMDB Votes vs Movie Revenue

C. Extraction of data from the-numbers.com

We scraped the site the-numbers.com to extract movie Budget of all the movies we were able to assemble from IMDB. To scrap the site we used scrapy, a free and open source web crawling and web scraping framework used to crawl websites and extract structured data from their pages, written in Python. The movie details were scrapped from the website and were stored on a separate csv file, which was later processed to remove invalid data. In addition to that, we used imdipy, imdbpy by imdbpy to collect budget from additional site such as fanpagelist.com, wikipidea, imdb.

D. Extraction of data from IMDB

We extracted the name of actors and directors from IMDB using IMDbPY. IMDbPY is a python package useful to retrieve and manage the data of the IMDb movie database about movies, people, characters and companies. Using the API given the movie name, we can search for all the actors and directors involved in the movie (i.e.: the data are fetched through the IMDb's web server http://akas.imdb.com and a SQL database, populated using the imdbpy2sql.py script)

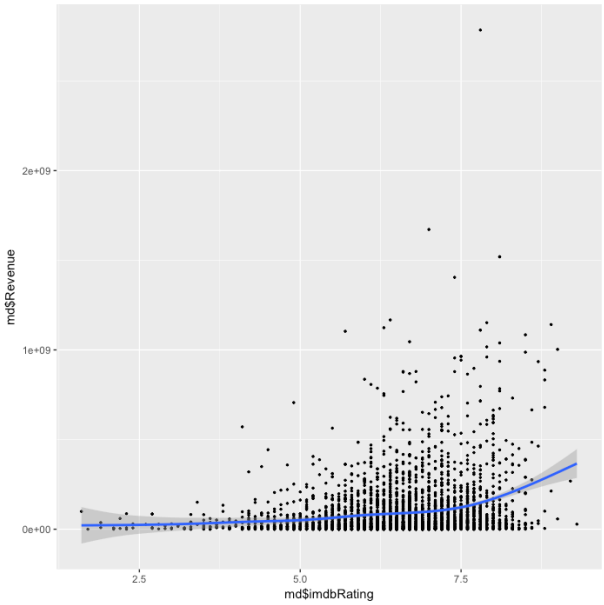

Fig. 8: IMDB Rating vs Movie Revenue

E. Understanding our Dataset

We have 3500 training instances and 13 features in our dataset. Some of the attributes that are present in our set is as follows: MovieDetail.csv: Training set of users
Set of Independent Attributes:

1) Title: Title of the Movie

2) imdbID: IMDB ID of the movie.
3) Actors-1: Lead Actor Name of the movie.
4) Actors-2: Second lead actor of the movie.
5) Actor1_Rev: Lead Actors last 3 movies revenue (averaged)
6) Actor2_Rev: Second lead Actors last 3 movies revenue (averaged)
7) Year: Year the movie was released
8) YT_Count: Youtube view counts for the movie (Summation of 1st three search result).
9) YT_Likes: Youtube Likes count for the movie (Aggregated to 1st three search result).
10) YT_Dislike: Youtube Dislikes count for the movie (Aggregated to 1st three search result).
11) IMDB Votes: No. of Votes received by the movie on IMDB site.
12) imdbRating: IMDB rating for the movie.
13) Budget: Budget of the movie (Scrapped from OMDB)
14) Revenue: Revenue earned by the movie (Scrapped from the-numbers.com)
15) Rated: MPAA rating for the movie.

The work flow for data collection System initially selected few movies from IMDB Database. The project work was broadly classified into following modules for application development.

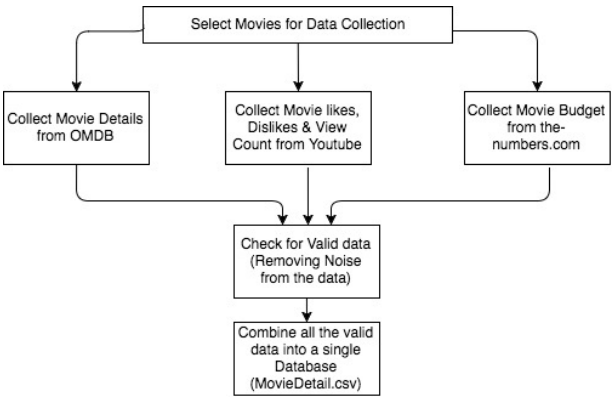

Fig. 9: Data Collection Flowchart

III. Methodologies

A. Data Analysis

The features were decided upon the fstatistic score on the models we created by the adding and removing indivisual parameters. This helped us to decide the effectiveness of these parameters and their potential impact on the revenue value.

Referring to Fig. 1, Fig. 2, Fig. 4, Fig. 6, Fig. 7, we see there are strong positive trends between Budget, Lead Actor Revenue, YouTube View Count, IMDB Votes with respect to the Revenue of the movies in our dataset which can be observed. Youtube Dislikes shows a strong negative trend. However, the IMDB Rating, Youtube Likes, Lead Actor do not contribute much to the final Revenue earned by the movie.

This can be also seen in the statistical analysis we have performed:

Attribute name	F Statistics	Adjusted R^2
IMDB Votes	844.3	0.3543
YT_Count	445.7	0.2244
YT_Likes	326.8	0.7149
Actor1_Rev	291.4	0.1589
Budget	245.9	0.1374
Actor2_Rev	220.7	0.125
IMDB Rating	85.91	0.05235
YT_Dislikes	4.033	0.0019

Before we take a combination of parameters to build a model, we will try to fit a model using revenue as the repsonse variable and each of our attributes individually. Upon performing an analysis on each fit, we observe the above listed F-statistic and Adjusted R^2 values. We see that IMDB votes, YT_Count, YT_Likes, Actor1_Rev have high F-statistic values. The same set of parameters also registered high Adjusted R^2 values. Interestigly, YT_Dislike has the lowest F-statistic and Adjusted R^2 value. But these are attributes which are taken individually and we are yet to see how they work in predicting the final revenue when these parameters work in combination.

So, we consider a set of attributes based on their F-statistic values, the adjusted R^2 values and the trends we see in the graphs. We make the following observations:

B. Model Building

Model Parameter	F Statistics	Adjusted R^2	Accuracy
Actor1.Rev, Actor2.Rev, YT_Count, YT_Likes, YT_Dislike, IMDB.Votes, imdbRating, Budget, Rated	241.3	0.5809	64.183
Actor1.Rev, YT_Count, YT_Dislike, IMDB.Votes, imdbRating, Budget, Rated	268	0.5809	66.61
Actor1.Rev, YT_Count, YT_Dislike, IMDB.Votes, Budget, Rated	283.7	0.5809	63.753
Actor1.Rev, YT_Count, YT_Dislike, IMDB.Votes, Budget	925.7	0.5715	74.928
Actor1.Rev, YT_Count, YT_Dislike, IMDB.Votes	670	0.4356	70.63
Actor1.Rev, YT_Count, YT_Dislike, Budget	760.5	0.4669	56.4471
Actor1.Rev, YT_Count, IMDB.Votes, Budget	1152	0.5704	69.914
Actor1.Rev, YT_Dislike, IMDB.Votes, Budget	1059	0.5496	68.338
YT_Count, YT_Dislike, IMDB.Votes, Budget	1148	0.5695	70.917

Based upon fstatistic score and adjusted rsquare values, we narrowed down on 2 linear regression models. We predicted the minimum possible movie potential using the linear regression models and used the box office performance criteria to verify the correctness. This criterion to check the performance gave us a performance accuracy in the range of 70%.

The linear regression model is used to predict the approximate figure of revenue. This helps to estimate the box office success type of the movie based on the profit it gains. This is very helpful for a studio to predict the movie before the movie is released or before the close of the first weekend. Without knowing the public perception also, the prediction can be done with around 56% accuracy based on the movie budget, the Rating awarded by the movie board, and the current Youtube view counts and reception of the movie trailer.

IV. Results

The results obtained are using a 3500 example data set. The train/test data split is a 70%/30% split. Even though it may be desirable to calculat the revenue potential, this information is hard to calculate based on the restricted data available and made public by the studios. Number of theatres the movies will be released in, number of countries and other such parameters are gradually disclosed by the studios and distributors and thus not available beforehand. This paramters will be needed to accurately predict the actual box office collections. But the prediction of whether the movie will be a blockbuster hit, or hit or average or flop will also help make decisions regarding promotions, campaigns and the other such marketing strategies.

Prediction	Results
Pre-Release Prediction	56.447
Post-Release Prediction with IMDB Data	74.928

V. Future Work

We plan on continuing to work on this project.The data set predicts the revenue based on the values collected from Youtube, IMDB as of now. In future we can collect data from Twitter, Google Trends (Hype factor) and Facebook.

We believe that this will increase the accuracy dramatically. We can search if the movie being predicted had any prequel, if one is available then the revenue of that movie will increase the accuracy. Other factors that can be considered are Critics rating/review, if the release date falls on a holiday, if the movie plot is an adaption from a famous novel or real life case.
In addition to performing simulations on a more representative data set, the results of this project can be improved by developing more robust methods of converting symbolic features into numerical values.

VI. Conclusion

The model we constructed helps to analyze the overall potential of the movie and predict its performance. If actual numbers are desired, more data related to marketing and distribution needs to be obtained. Whereas, using the social parameters from various social networks allows us to build a rough prediction model useful to get the revenue and profit estimates.

VII. Acknowledgments

The authors would like to thank Professor Anjum Chidda and Mr Vatsal Patel for their support and advice on this project.

References

[1] Webpage crawling data scraper API [Online]. Available: https://scrapy.org/
[2] Python Panda Package for Data Analysis [Online]. Available: http://pandas.pydata.org/
[3] IMDB Movie Database API [Online]. Available: http://www.omdbapi.com/
[4] IMDB Movie Details Page [Online]. Available: http://www.imdb.com/
[5] Budget of Already released Movies [Online]. Available: http://www.the-numbers.com/
[6] Movie Insight [Online]. Available: http://fanpagelist.com/
[7] Youtube Developer's Data API [Online]. Available: https://developers.google.com/youtube/v3/
[8] Wikipedia Movie page [Online]. Available: https://en.wikipedia.org/wiki/Avatar_(2009_film)
[9] Lyric Doshi, Using Sentiment and Social Network Analyses to Predict Opening-Movie Box-Office Success.
[10] Michael T. Lash and Kang Zhao, Early Predictions of Movie Success: the Who, What, and When of Profitability
[11] Wenbin Zhang and Steven Skiena, Improving Movie Gross Prediction Through News Analysis
[12] Matt Vitelli, Predicting Box Office Revenue for Movies
[13] Jason van der Merwe and Bridge Eimon, Predicting Movie Box Office Gross Shivam Mevawala and Sharang Phadke, BoxOffice: Machine Learning Methods for Predicting Audience Film Ratings
[14] L. Barry, Predicting Success of Theatrical Movies: An EmpiricalStudy, Journal of Popular Culture, 2004.
[15] Tom Mitchell, "Machine Learning - Tom Mitchell".
[16] Christopher Bishop, "Pattern Recognition and Machine Learning".
[17] IMDbPY, Python package [Online]. Available: http://imdbpy.sourceforge.net/index.html