Predicting a Film's Rating Utilizing Machine Learning

Joel Martinez
Texas State University

Jose Garcia
Texas State University

Kameron Bush
Texas State University

April 1, 2019

Abstract

We present a program that will utilize existing data on films that have been released in order to predict the rating of a film that has yet to be released. This prediction will be given to the film industries that created the film in order to give them a chance to improve the film if needed. The factors being used are IMDB rating, budget, USA gross, USA opening weekend gross, tomatometer, metascore, and worldwide gross. We used using linear regression as the model for the project and it will be optimized using the k-fold method.

1 Problem Description

When a film industry decides to create a film, the creators have to factor in many attributes that are important to how the public will interpret and perceive the film. Some of these factors are the companys film budget, the film critic ratings, and how much the film makes on its first weekend.
This public interpretation is very important to the film company because how the public perceives the film ultimately determines how many people go and see the film. The amount of people who go see the movie ultimately determines if the film company makes a profit off of the movie, i.e. if it is a hit or a flop.
The fate of the film cannot be determined until after the film is released to the public making it too late to make any changes if it fails. Our program is aimed to fix this problem by analyzing factors of previous films in order to predict the audience rating the new film will receive. This audience rating will give the film company a general idea of how well their film will do before they release the movie. Based on the audience rating presented, the film company will be able to fix their film before release so they can maybe save the film from flopping.

2 Survey

To see how much of an impact this project would have with the general population, we asked 50 students their opinion.

Q: Do you believe this project will have an impact on which movies you will watch based on the rating generated?

37 students said it would have an impact.
9 students said it would not impact them.
4 students said they did not know.

3 Plan

March 8th - Understand the Problem in-depth
- Our group has got a now gotten a better understanding of how we can attack this problem. Our attributes will be able to help us use our model to estimate the probability of success (IMBD Rating, Budget, USA Gross, USA Opening Weekend Gross, Tomatometer, Metascore, Worldwide Gross).

March 14th - Develop a data base for the code
- We first created a database where we stored our x attributes and our y audience rating (1.0-10.0) in excel.

March 25th - Complete a base algorithm for Project
- Our model first takes the data from the database (excel), then splits our data to a X and Y dataset. Our model will consist of a train/test split of the X and Y dataset. We used Linear Regression to find a relation between the X attributes and the rating, we then proceeded to use the K-Fold method to optimize our model.

March 29th - Complete Intermediate Project Report
- Done and done.

April 10th - Research ways to make algorithm more accurate
- Yet to be done.

April 20th - Edit algorithm to be more effective and more precise
- Yet to be done.

April 29th - Complete Final Project Presentation
- Yet to be done.

May 4th - Complete Final Project Report
- Yet to be done.

4 Data

The data set contains seven attributes (or features, denoted by X1. . . X8) and one response (or outcomes, denoted by y1). The aim is to use these seven features to predict the outcome.
We specifically choose these seven features because the information is easy to find and open to the public. Some other attributes that we considered for the data was harder to come across on the internet and left many holes in our data. This would have required us to have to clean our data beforehand. With the attributes we are using there is no need to clean the data because we handpicked each movie individually to contain all or most of the attributes.
Data was obtained from the Rotten Tomatoes website which provided us with the Tomatometer (Rotten Tomatoe's score of the movie), and the audience score (the value that we intend to predict.) We also utilized the IMDb website to obtain the gross USA (the amount of money that was made in the USA), the gross opening weekend USA (the amount of money that was made in the first week that the movie was released to the public), worldwide gross (the total amount of money made around the globe), budget (the amount of money the studio spent to create the movie), metascore (the IMDb user's score of the film), and IMDb rating (the score that IMDb critics gave the film.)

Specifically:

X1 Gross USA

X2 Gross Opening Weekend USA

X3 Worldwide Gross

X4 Budget

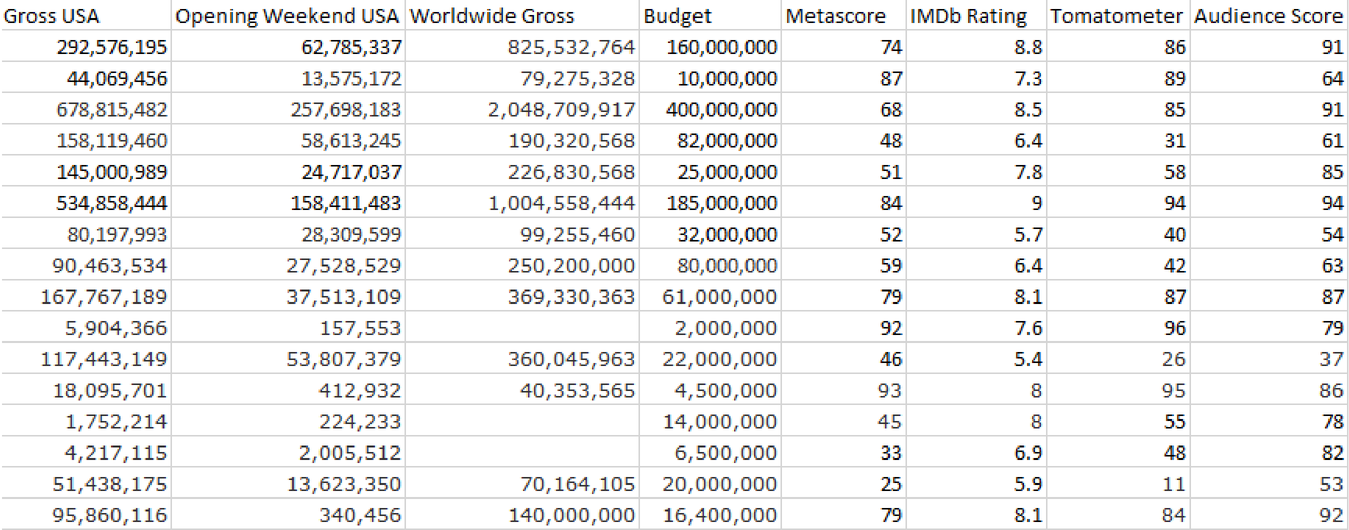
X5 Metascore

X6 IMDb rating

X7 Tomatometer

y1 Audience Score

Below is a screenshot of the first few lines in the dataset that we are utilizing for our model. As you can see, most attributes are filled in and few are null. We intend to expand the size of of the dataset as we continue to work on the project. Much of the data is gathered by hand and requires a large amount of time to find.



Gross USA	Opening Weekend USA	Worldwide Gross	Budget	Metascore	IMDb Rating	Tomatometer	Audience Score
292,576,195	62,785,337	825,532,764	160,000,000	74	8.8	86	91
44,069,456	13,575,172	79,275,328	10,000,000	87	7.3	89	64
678,815,482	257,698,183	2,048,709,917	400,000,000	68	8.5	85	91
158,119,460	58,613,245	190,320,568	82,000,000	48	6.4	31	61
145,000,989	24,717,037	226,830,568	25,000,000	51	7.8	58	85
534,858,444	158,411,483	1,004,558,444	185,000,000	84	9	94	94
80,197,993	28,309,599	99,255,460	32,000,000	52	5.7	40	54
90,463,534	27,528,529	250,200,000	80,000,000	59	6.4	42	63
167,767,189	37,513,109	369,330,363	61,000,000	79	8.1	87	87
5,904,366	157,553		2,000,000	92	7.6	96	79
117,443,149	53,807,379	360,045,963	22,000,000	46	5.4	26	37
18,095,701	412,932	40,353,565	4,500,000	93	8	95	86
1,752,214	224,233		14,000,000	45	8	55	78
4,217,115	2,005,512		6,500,000	33	6.9	48	82
51,438,175	13,623,350	70,164,105	20,000,000	25	5.9	11	53
95,860,116	340,456	140,000,000	16,400,000	79	8.1	84	92

References

[1] Quader, Nahid & Gani, Md. & Chaki, Dipankar & Ali, Md. *A Machine Learning Approach to Predict Movie Box-Office Success.* Reading, Bangladesh, 2018.

[2] Sharda, Ramesh & Delen, Dursun *Predicting box-office success of motion pictures with neural networks.* Expert Systems with Applications. 30. 243-254. 10.1016/j.eswa.2005.07.018.

[3] B. R. Litman & H. Ahn *Predicting financial success of motion pictures.* In B. R. Litman (Ed.), the motion picture mega-industry. Boston, MA: Allyn & Bacon Publishing, Inc. (1998)

[4] M.H Latif & H. Afzal *Prediction of Movies popularity Using Machine Learning Techniques.* National University of Sciences and technology, H-12, ISB, Pakistan.