IST 3420 Group Project

IMDb Exploratory Analysis

Adam Forestier, Parika Gupta,

Brandon Jones, Nathaniel Williams

Last update on: September 7, 2017

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Version** | **Description** | **Author** |
| <dd/mmm/yy> | <x.x> | <details> | <name> |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

[**Introduction**](#_30j0zll) **4**

[**Data Source and Collection**](#_1fob9te) **5**

[**Cleanse and Transform Data**](#_3znysh7) **7**

[**Analyze Data**](#_2et92p0) **8**

[**Summarize Findings**](#_tyjcwt) **9**

# **Introduction**

The primary goal of our project is to determine if we can closely predict movie ratings before critic ratings are released and the films are given an average user score on IMDb. Being able to determine whether or not a movie will be a quality release without critic ratings could address a number of business problems that surround the financing and marketing strategies for different types of movies. Film studios and movie theatres could use these insights to budget accordingly and properly advertise movies to maximize popularity, ratings, and profits. The dataset we will be using contains 5,043 movies, and 28 variables related to the movie.

The primary research questions we wish to ask revolve around predicting a movie’s IMDb score. One of which, is will the number of faces in a movie poster correlate with the movie’s rating? Can the multiple actor variables and the director variable be useful in predicting a movie’s IMDb score? Does a movie’s parental guidance rating (meaning R, PG-13, etc.) play into how well a movie scores? Does a movie’s budget help determine a movie’s critical success? Do certain genres tend to receive higher critic scores?

The secondary questions we are interested in solving revolve around being able to determine a movie’s revenue based on the following criteria: Do different genres of movie play into a film’s revenue total? Do critic scores determine how much a movie grosses? Does the number of faces on a poster play into how much a movie grosses? How much do popular actors and directors determine a movie’s financial success?

Tertiary questions we wish to investigate are general questions about how variables relate to each other. We wish to find out if different nation’s films tend to have different duration times. We would like to see how gross compares to budget and determine what budget expenditure typically leads to the best returns ratio-wise. We also would like to determine what key plot words are used most commonly, and if they lead to commercial and critical success.

# **Data Source and Collection**

Our data was retrieved as a CSV file from Kaggle: <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

The data was retrieved by scraping 28 values for 5043 movies from the IMDb website and by using facial recognition software run on 4906 posters. The movies’ release dates span over 100 years and the movies originate from 66 different countries. The variables of the dataset are as follows:

* movie\_title
* color num\_critic\_for\_reviews
* movie\_facebook\_likes
* duration
* director\_name
* director\_facebook\_likes
* actor\_3\_name
* actor\_3\_facebook\_likes
* actor\_2\_name
* actor\_2\_facebook\_likes
* actor\_1\_name
* actor\_1\_facebook\_likes
* gross
* genres
* num\_voted\_users
* cast\_total\_facebook\_likes
* facenumber\_in\_poster
* plot\_keywords
* movie\_imdb\_link num\_user\_for\_reviews
* language
* country
* content\_rating
* budget
* title\_year
* imdb\_score
* aspect\_ratio

Much of our data will need to be cleansed. The variable “director\_facebook\_likes” is inaccurate in many instances. Foreign films using currencies different from USD may have a misrepresented gross revenue, because the values aren’t equivalent with foreign currencies. This can be addressed by transforming these values by an exchange rate. We would also like to retrieve the release date for each film, since that variable was not included in the original dataset.

The goal of our project is to find connections between variables and the film’s IMDb score. We believe that the key variables are the following: facenumber\_in\_poster, director\_name, actor\_1\_name, actor\_2\_name, and actor\_3\_name. Our hypothesis is that these variables will be crucial in predicting a movie’s overall critic score. We also hypothesize that budget, gross, actors, directors, genre and content rating will play into how much a movie grosses in revenue.

# **Cleanse and Transform Data**

< Extract and transform potential variables from the data source(s). Cleanse your dataset(s).

Add sub sections if you need.>

# **Analyze Data**

<Analyze your dataset using various approaches. Add sub sections if you need.>

# **Summarize Findings**

<Summarize insights you obtained from the data analysis. Propose suggestions for business value creation and performance improvement.>