Data collection and preparation

For making our data set, we used two sources. The first one is Box office Mojo website, and the second is IMDB. We used the python library beautiful soup to read the HTML webpages and utilize regular expressions to extract specific string pieces. In our study, we focused on the US movie market between 2010 and 2019. So we only pull movies’ information which has been released during this period. To gather the required information, we took the following steps for each film:

1. Go to boxofficemojo.com
2. Go to each movie page and read the page and extract the required information
3. Go to IMDB.com
4. Read the movie IMDB page and extract the information

In fig 1, you can see the schematic picture of this procedure.

Dataset description

Our data consists of 8403 movie information that has been released in the US from 2010 to 2019. This dataset has 27 columns (features) which we describe below:

The boxofficemojo information:

IMDB ID: A unique combination for each movie on the IMDB website

Mojo ID: A unique combination for each movie boxofficemojo website

Title: The original title of the movie

Genres: The Genre(s) of the movie

Year: The year which the film went on screen

Domestic Gross ($): total income of the movie in the US in dollars

Worldwide Gross ($): net income of the movie in the world

Opening ($): Opening weekend gross of the movie

Budget ($): the amount mony that is spent to make the movie

Opening Theaters: number of theaters that show the movie when it is initially released

MPAA: the age rating of the movie

In Release (Days): number of days that the movie has been on the screen

Widest Release (Days): the most significant number of theaters that showed the movie

The IMDB information:

Stars: The name of the five most important actors in the movie

Director(s): the director(s) of the movie

Writer(s): the writer(s) of the movie

Producer(s): the producer(s) of the movie

Runtime: duration of the film in minutes

IMDB score: the IMDB score of the movie

IMDB votes: number of the people who vote for that movie on IMDB

Metascore: the score of the movie on the Metacritic website

Meta users: the composition of the people (regular user or critic) who wrote a review about the movie

Country: the country that made the movie

Language: The original language of the movie

Distributor: The company which distributed the movie

Plot outline: a short description of the movie on IMDB

Data processing

As you can see, our data contains both numerical and non-numerical values. For using this data in machine learning algorithms, we should convert non-numerical values to numerical ones.

First, we begin with the release date of the movies. The date is made of a number and a word like ’18, Jan’. To convert this data to only numerical values, we first separate the number and then attribute a number from 1 (January) to 12 (December). So the release date column becomes two separate columns, namely ‘Day’ and ‘Month.’

For converting Genre columns to numerical values, we first find all genres that present in our dataset. We have 26 categories in total: 'Action', 'Adult', 'Adventure', 'Animation', 'Biography', 'Comedy', 'Crime', 'Documentary', 'Drama', 'Family', 'Fantasy', 'Film-Noir', 'History', 'Horror', 'Music', 'Musical', 'Mystery', 'News', 'Reality-TV', 'Romance', 'Sci-Fi', 'Short', 'Sport', 'Thriller', 'War', 'Western'. We first sort these categories alphabetically and number them from zero to 25, then assigned each category a unit vector. All of the elements are zeros except the index, which is the same as that category’s position in the category list.

We do the same for MPAA column, except this time we have 9 categories: 'G', 'M/PG', 'NC-17', 'Not Rated', 'PG', 'PG-13', 'R', 'TV-PG', 'Unrated'.

Exploring the data

It is always good to see how our data looks like before we feed it to our machine learning algorithms. In fig you can see the correlation between differnet features.

As you can see,

In fig you see the histogram of worldwide gross. Because most movie sales are low, it is better to plot the histogram of log of worldwide gross.

In fig and you see the scatter plot of worldwide gross vs budget of movies and language vs. worldwide gross box plot, respectively.

Fig shows the most frequent words in movie titles. The bigger the word is, the higher frequency that word appeared in the title of the movies.