Project title: Movie Success Prediction

Team Members: Yonatan Cipriani & Hamza Khanani

Intro + Motivation:

For this project, we will be analyzing movies from the past and see what factors have established those movies as flops, hits, or blockbuster. Millions of dollars are being poured into Hollywood and the movie industry, therefore it is important to identify what makes a movie flourish at the box office. Successfully predicting the fortunes of an impending movie release would be beneficial to everyone who was involved in the movie creation process. If a movie is considered a flop then that means that is not profitable during its time in theaters. Once a movie is predicted to be unprofitable there are several options that may be taken. This would either lead to shifting strategies and creativing marketing decisions or the movie may just be axed. It is important for movie executives to deliver movies that can maximize their profits at the box office. By having growing, historical data available machine learning techniques can be applied in order for organizations to have a better chance of identifying what is profitable and what is a risk. Our project can also be very useful for the viewers who are not willing to waste their money and time at the theatre as the algorithm we are implementing will give them a rough idea of what to expect from the movie.

Objectives:

We will be using a dataset that we obtained from Data.world. The dataset contains many features such as movie budget, directors, content rating, title year, and so on. We will need to compact that list use the metrics involved to predict our main feature, Internet movie database (IMBD) score. We will be dividing our rating which we predicted into several intervals. An IMBD score from 1-5 would deem the movie as a flop. A rating of 5-8.5 would make the movie a hit. Finally, a score of 8.5-10 would consider the movie to be a blockbuster. By accurately predicting what category a movie will fall under, we will be achieving our goal. Once movie stakeholders have this prediction, they will be able to take the necessary steps to make it more profitable or axe the movie if needed.

Methods:

In order to implement this project we will need to explore different methods that will help us achieve the highest accuracy of a move prediction. As of now, we are planning to use a decision tree based on historical data of past movies which have in our dataset. Another possible method to use would be implementing our own weighted average function algorithm. The k-fold cross validation method may also be beneficial to use since it provides an average error that we can use to check the accuracy. Finally, we will try to implement a kmeans clustering algorithm that will cluster the movies into three different clusters, a cluster belonging to flop movies, hit movies and blockbusters.We will divide the workload of this project by having each team member explore a different method and analyze the outcome. In the end, we will combine the knowledge we gained and draw conclusions.

Literature Search:

There are many machine learning algorithms that have been implemented for early prediction of a movie's success. We got a chance to go through some literature already present and it looked like that genre was one of the biggest factors of the movie being a flop, hit or a blockbuster. One of the other factors which highly affected the rating were the main actors. Most of the well known actors have a reputation because of their past work and viewers usually give a good rating if the actor was up to the mark. The literature we read has also helped us consider the weightage we will be giving to some specific features which might lead to a better accuracy.
Each implementation chooses specific features and defines the "success" a little differently. We will analyze some of the implementations results with ours and analyze the similarities and differences. Also, each of us will look at the other methods that were used for the prediction and will possibly try to implement our own version of those methods if has better results (accuracy) than ours. The literature present also gave us a high level overview of the techniques used in the past and how successful those models were.

Timeline and Expectations:

We will begin to implement the different methods for implementing the movie prediction after our last homework assignment is due (November 6). Hopefully, two weeks will give us enough time to actually have working code for predicting the movie success (November 20). This will give us a week to check the our accuracy and analyze the methods and data we have completed (November 27). This will leave us with a week to draw conclusions from the data and create the project video presentation (December 4). Finally, we can write out the project report which will be due the following week (December 11). Overall we are expecting our algorithm to be very accurate, we'll be testing it using the existing data set we have as it already has the IMDB rating for all the movies. We will also be using random movies which have already been released and predict an IMDB rating using our algorithm to compare it with its actual rating.