CSE 519 PROJECT PROPOSAL

Which movies endure and why?

Objective:

To build a model that predicts the current popularity of a movie as a function of variables known at its time of release.

Background Research:

The IMDb dataset is used to collect data regarding the movie, cast ,box office opening etc.Kaggle Academy awards dataset is to be integrated with IMDb data to see if we can correlate awards with how long a film will endure and why .Google Trends API offers a layer for Google Trends Data, which can be used to retrieve keyword popularity data by location and date.

There has been some research done in the past using movie dataset to predict their popularity.

Paper[1] examines the genre trends in the top 50 grossing films the US box office each year from 1991 to 2010.Focuses on the frequency and rank of different genres and the box office gross.The trend in the time series of the frequency with which a genre occurs in each year is described by linear ordinary least squares(OLS) model, y=at+b where the independent variable t is the time of release.Reduced major axis(RMA) regression was used to predict the relationship between the ranks of a film's opening and total gross.Concludes that the range of genres for highest grossing films has become narrower over the past twenty years and that different genres are characterized by different release patterns.

Paper[2] uses hybrid visually-driven features which represent the movie content to predict movie popularity and average rating of the movie.The movie was split into shots and a frame is selected as representative of each shot.From this frame the features were extracted and then aggregated to build an individual feature vector that is descriptive of the movie.These features are then fed to a gradient booster model for training nd prediction.It was identified that popularity and average rating were positively correlated and naturalness ,which is one of the visual features was highly correlated with average rating.

Paper[3] built a naive bayes model that uses character and plot based user-identified trope labels to predict audience rating for movies.After comparing the accuracies of various models they conclude that the narrative tropes used to create a movie are not very predictive of the final rating or quality of the movie.

Paper[4] predicts movie box office performance using followers count on twitter and sentiment analysis of YouTube viewers' comments.The predictions were labeled as three classes,Hit,Neutral and Flop using K-Means clustering.The following patterns were identified: a new movie in non-popular genre and an actor with low popularity resulted in a flop ,popularity of leading actress is crucial to the success of the film.

Paper[5] evaluates if the success of a film can be predicted well with data obtained through conventional methods such as web databases or with data obtained from social media such as Twitter,Youtube.The results demonstrate that the sentiments harnessed from social media can predict the success with more accuracy than that of using conventional features.A blend of both outperformed the existing approaches.

Paper[6] predicts song popularity based on it’s audio features and metadata.Different classification and regression algorithms were evaluated based on their ability to predict popularity and determined the types of features that held the most predictive power.The popularity metric used for training was obtained from Echo Nest.The most influential features were artist popularity ,loudness, year and genre tags.

Dataset:

Following datasets and sources have been identified for use.
IMDb Dataset.
Box Office Mojo
TMDb API
Kaggle ‘Academy Awards’ Dataset.
Rotten Tomatoes.
Twitter/FB API
YouTube Data API

The IMDb Dataset probably is the most comprehensive Internet Movie DataBase of information related to films, television programs including cast, production crew and personal biographies, plot summaries, trivia, fan and critical reviews. The IMDb provides you with the STARmeter, MOVIEmeter, and COMPANYmeter which indicate the level of public awareness and/or interest in the title, person or company.

Analyzing deeper into the popularity of the movie, we strongly believe that the Academy awards dataset from kaggle would give us interesting insights about whether the most popular movies win Oscars and if Oscars affect the popularity of the movies henceforth. The data here contains the Academy award nominees and winners from 1927 until today for various categories offered. Therefore, analysis for the Oscars shouldn’t be a problem.

The TMDb API provides us with the popularity score of the movies which helps us to understand why some movies are more popular over others. Also, it provides us with the keywords for each particular movie allowing us to broaden our understanding of how certain plots have stood the test of time.

To understand the popularity of the movie, we intend to collect the box office sales of the movie from the Box Office Mojo for the first two weekends and the total gross collections of the movie which would assist us in determining how the first weekend collection has enhanced or hampered the expectations on a particular movie. This also gives us the total number of screenings for the movie each week.

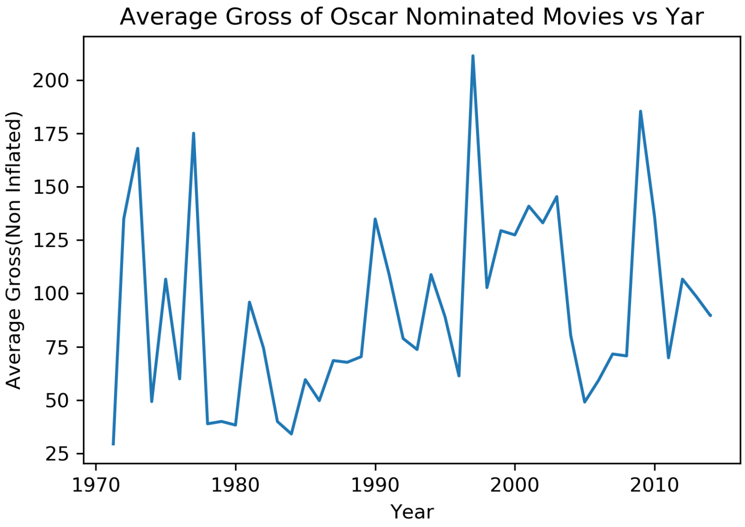
The Twitter API gives us the twitter analytics like the average tweet performance, engagement rate, content performance, brand follower growth, hashtag performance, Impressions etc for the various movie pages as well for celebrities on twitter.

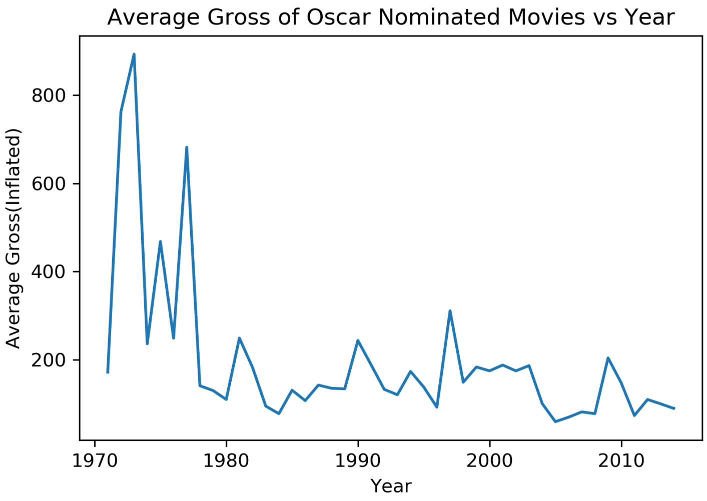
Moreover, to get grasp of the popular movies of the yesteryears, we plan to scrape the data of the DVD sales, VHS sales, from The Numbers to give us better insight into the overall popularity. Further, the rotten tomatoes ratings are used to understand how the popularity, success, failure of the movies at the box office are affected by the Tomatometer after its first weekend.

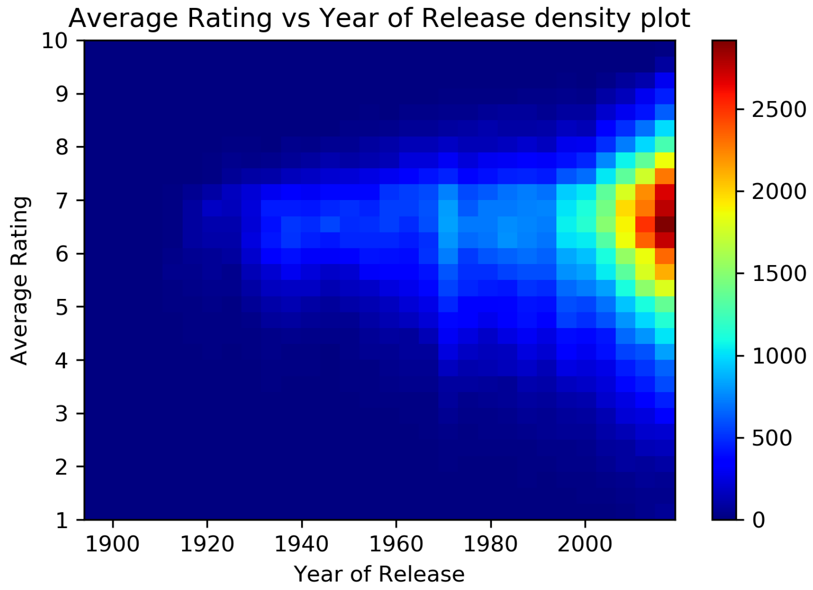
Apart from classifying movies based on genre, we plan to classify them based on the seven basic plots: : 1.Overcoming the Monster, 2.Rags to Riches, 3.The Quest, 4.Voyage and Return, 5.Rebirth, 6.Comedy and 7.Tragedy. After this, we plan to check if certain plots have become outdated over time like how certain genres of movies have become.

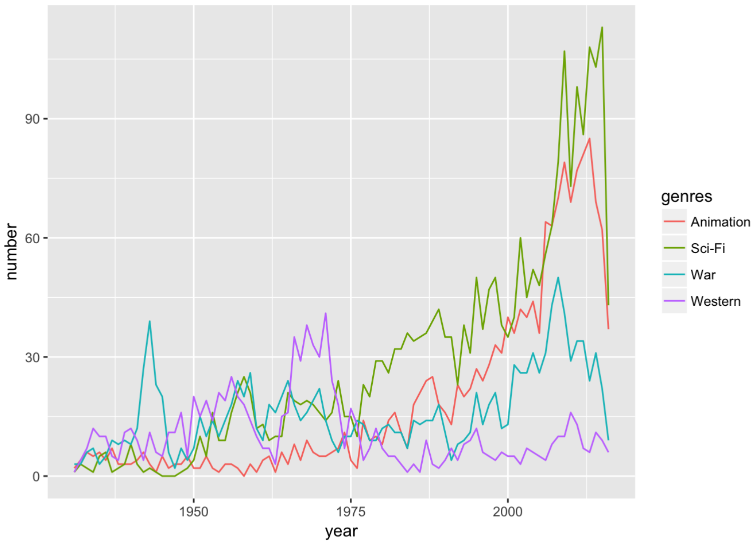
Exploratory Analysis:

● The average money made in the box office by movies nominated for the Best Picture in Oscars between 1970 to present is plotted and can see that while the non inflated gross has been increasing over the years, plotting the box office gross taking inflation into account shows that it has actually been decreasing over the years, showing the trend that popular movies which make money are usually the movies which win Oscar awards.


Average Gross of Oscar Nominated Movies vs Yar


Average Gross of Oscar Nominated Movies vs Year


Average Rating vs Year of Release density plot

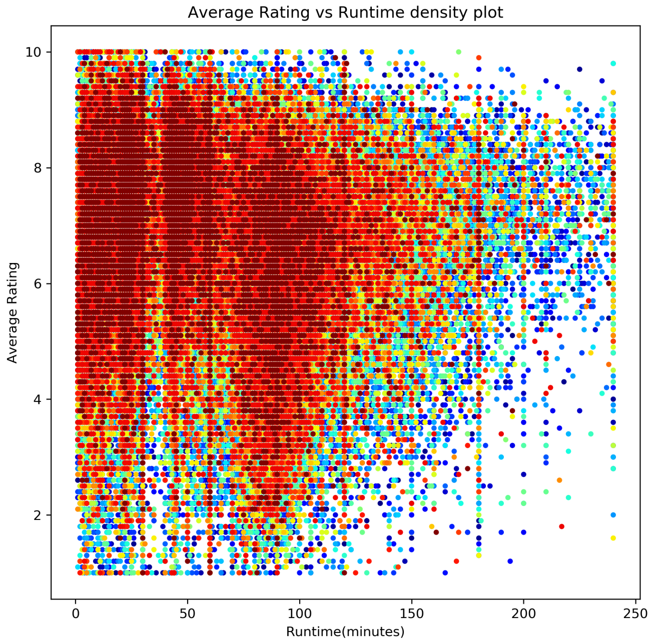


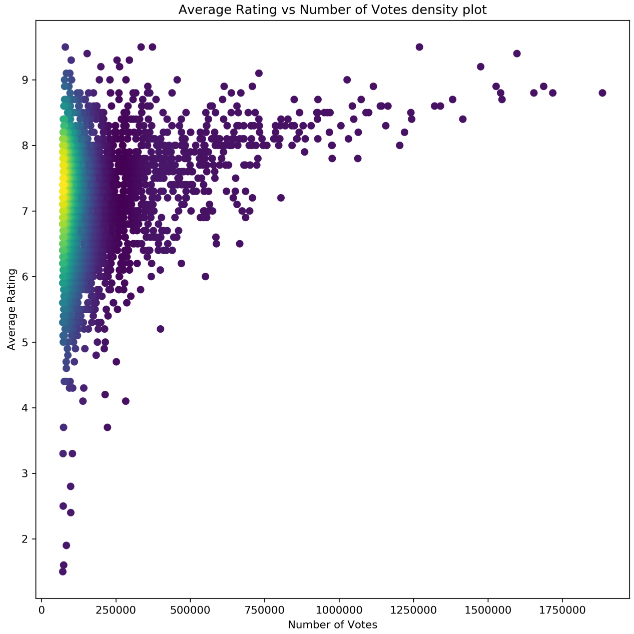
●The number of films made in the genre ‘Western’ has decreased over the years while it has increased in the genre ‘Sci-Fi’

●A 2D histogram is used to plot the intensities of average ratings vs year of release and specify the bin size to show the density of the scatter plot. The plot shows that the number of ratings given for movies after 2000 is more compared to before and it lies between 5 and 8.

● A scatter plot between the average rating for a movie and number of votes is plotted.As the number of data points is very large, the density plot is shown as a heat map. We can conclude that most movies have number of votes less than 10000 and average rating between 6 and 8.

●A scatter plot between the average rating for a movie and the runtime in minutes is shown below.Runtime unto 4 hours is only considered and TV shows and shorts are removed. We see that there are a lot of movies with average rating between 6 and 8 and runtime close to 90 minutes.





Challenges:

●The datasets does not provide us with the popularity of the movies. Writing a metric for the popularity index pose us with many sub challenges.

●The metric can be designed by the number of times it has been searched on IMDb but IMDb does not provide us with an API to extract that data. Though the TMDb Kaggle data gives us with the popularity index, it’s only for 5000 movies.

●Even though Google Trends gives us the number of times a particular title has been queried, aggregating the data is very challenging and there is data available only from 2003.

●The social media views on Twitter/YouTube/Facebook give you insights about the popularity of current movies but they are heavily skewed against the movies of the past, say a century ago.

●The popularity of the old movies can be determined by the number of VHS, DVD, number of days it had lasted in theaters and the number of theaters it had run but the data isn't available.

●There are 793816 films that are recognized as Short, but around 14384 actually don't belong in this genre as the time duration is greater than 45 minutes. The wrong data is because the films share similar names.

●The old movies were rated 1-4 while the current movies are rated between 1-10 and there is a need to scale it. The scaling might not produce values comparable to the current time.

● The comparison of old movies- rated by DVD sales- and the new movies- rated by social media views- gives biased results.

●For all the box office sales and the theatre numbers, the values are affected by inflation.

●While analyzing the IMDb data, we found that many of the titles had contained a blank space at the end of the name which makes it harder to analyze the data.

●Data was lost while merging two data sets because of the mismatch in the names. For eg, Birdman was named as 'Birdman or An unexpected Virtue of Ignorance' in one of the datasets.

●The numeric columns had values from another column leading to a datatype mismatch.

●Incomplete data for the old movies and it's difficult to impute them and therefore, popularity of the movie can't be defined easily.

●There are many columns with string data holding many attributes within it and handling them is tedious.

Approach:

●Primarily, we would like to compute the popularity metric by considering various parameters like the actor, his previous film success rate, the director and his success rate, the budget, the production company, the run time, the genre, the keywords and the year of release to find out the popularity metric. The budgets are scaled by the CPI index. The run time is scaled considering the year of release.

●Then, we calculate how popular the movie is by considering the box office sales, the theatre numbers, the number of times it has been quoted on social media, any references of it used in the movies later. For old movies, we plan to obtain the data of the DVD sales from the Numbers. We plan to scale to match with the current social media activity. This gives us how much more/less popular the movie is. Also, we plan to use the popularity score from the Kaggle TMDb data to compare our metric.

●For the categorical data, we plan to filter the top 10% of the movies based on the ratings for each year. Then define a function which would give us the top five high frequency parameters. Then, we repeat the same procedure to filter the bottom 10% of the movies.

●For continuous parameters like budget, revenue, views, hits, likes , run time, we divide it into various levels and then find the corresponding rating for each level. Then we intend to plot them. Doing this, we report the 10 films with the dramatically highest/lowest performance and analyze how various features are correlated to the performance of the film.

●Also, we group by genre and pull out their average ratings for two different decades. Then we plan to do a significance test (T-test) within each genre and between genres, to see if there are significant differences in the change of ratings throughout these 10 years.

Validation:

We plan to use Google Trends which analyzes the popularity of top search queries in Google Search across various regions and languages.The website uses graphs to compare the search volumes of different queries over time.If there has been a consistent amount of searches regarding a movie over a period of time then it can be deemed popular.Another way to validate the prediction is to use the trending movies feature on Netflix.

Reference:

1. Using Crowd-source based features from social media and Conventional features to predict the movies popularity by M.Ahmed,M.Jahangir,Dr.Hammad Afzal,Dr. Awais Majeed,Dr.Imran Siddiqi
2. Predicting movie success from tweets by Rose Catherine, Sneha Chaudhari
3. SOCIAL MEDIA,TRADITIONAL MEDIA AND MUSIC SALES by Sanjeev Dewan , Jui Ramaprasad
4. K. Apala, M. Jose, S. Motnam, C.-C. Chan, K. Liszka, and F. de Gregorio. Prediction of movies box office performance using social media In Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, pages 1209–1214, Aug 2013
5. Predicting Movie Popularity and Ratings with Visual Features by Farshad B. Moghaddam, Mehdi Elahi, Reza Hosseini, Christoph Trattner ,Marko Tkalčič
6. Predicting the movie popularity using user-identified troops by Amy Xu,Dennis Jeong.