Actor - Genre Classification and IMDB score Prediction

Kiran Karpurapu (kvk229) and Gargi Singh Chattwal (gsc326)

Describing the problem:

a) What is the problem?

The main idea of this project is to, for every actor/actress, assign weights for his strengths in about 10 selected genres. With this kind of data, the actor/actress knows what kind of movies are people expecting from him/her.
As a secondary idea, we wanted to work on predicting the IMDB score of a movie based on a few variables which are dealt in the succeeding sections by learning from data from history. We also want to investigate on the effectiveness of various machine learning algorithms on predicting the IMDB score for a movie.

b) What types of problems are these?

The first part of the problem involves multi-class classification with a probabilistic approach for the class labels and the second part is a typical logistic regression type of problem. To achieve the stated goals, there is a lot of number crunching involved.

Background:

a) Initial motivation:

Considering the scale of market/demand available for the movie industry, it is very important for actors to stay alive in the competition by making pleasing movies. Not all actors can work well in all the genres and the audience expect the best performance from their favourite actor always. So we thought it is always beneficial for an actor to pick movies from the genre where he is good at and this project is aimed in assisting an actor in picking his next genre.

b) Learning about the background:

There is no particular domain expert that we will involve with but there are multiple research teams doing similar kinds of projects in both academia and the industry.
People have written research papers from which we were inspired to choose this topic.
Few related links:

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=7&ved=0ahUKEwjxmODY3eXPAhWMcz4KHdagBrAQFghOMAY&url=http%3A%2F%2Fcs229.stanford.edu%2Fproj2013%2Fcocuzzowu-hitorflop.pdf&usg=AFQjCNFaXKlewK5TYu3XRsleFK-EHzAMKw&sig2=EUdWR2d1ElHe3rM30i55NA&bvm=bv.135974163,d.cWw&cad=rja

https://chrisarcand.com/movie-projections-using-relational-databases-and-classification/

Describing the data:

a) Kind of data:

We will be using a huge CSV file that was obtained by someone scraping the IMDB website for movie related data. Data contains info of about 5000 movies over 60 years duration.

b) Data Source:

We obtained the data from the website: **https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset.**

c) Features of the data:

$ color	: Factor w/ 3 levels
$ director_name	: Factor w/ 2399 levels
$ num_critic_for_reviews	: int
$ duration	: int
$ director_facebook_likes	: int
$ actor_3_facebook_likes	: int
$ actor_2_name	: Factor w/ 3033 levels
$ actor_1_facebook_likes	: int
$ gross	: int
$ genres	: Factor w/ 914 levels
$ actor_1_name	: Factor w/ 2098 levels
$ movie_title	: Factor w/ 4917 levels
$ num_voted_users	: int
$ cast_total_facebook_likes	: int
$ actor_3_name	: Factor w/ 3522 levels
$ facenumber_in_poster	: int
$ plot_keywords	: Factor w/ 4761 levels
$ movie_imdb_link	: Factor w/ 4919 levels
$ num_user_for_reviews	: int
$ language	: Factor w/ 48 levels
$ country	: Factor w/ 66 levels
$ content_rating	: Factor w/ 19 levels
$ budget	: num
$ title_year	: int
$ actor_2_facebook_likes	: int
$ imdb_score	: num
$ aspect_ratio	: num
$ movie_facebook_likes	: int

Title: "The Chronicles of Narnia: The Lion, the Witch and the Wardrobe",
Year: "2005",
Rated: "PG",
Released: "09 Dec 2005",
Runtime: "143 min",
Genre: "Adventure, Family, Fantasy",
Director: "Andrew Adamson",
Writer: "Ann Peacock (screenplay), Andrew Adamson (screenplay), Christopher Markus (screenplay), Stephen McFeely (screenplay), C.S. Lewis (book)",
Actors: "Georgie Henley, Skandar Keynes, William Moseley, Anna Popplewell",
Plot: "Four kids travel through a wardrobe to the land of Narnia and learn of their destiny to free it with the guidance of a mystical lion.",
Language: "English, German",
Country: "USA, UK",
Awards: "Won 1 Oscar. Another 14 wins & 44 nominations.",
Poster: "https://images-na.ssl-images-amazon.com/images/M/MV5BMTc0NTUwMTU5OV5BMl5BanBnXkFtZTcwNjAwNzQzMw@@._V1_SX300
Metascore: "75",
imdbRating: "6.9",
imdbVotes: "289,869",
imdbID: "tt0363771",
Type: "movie",
tomatoMeter: "76",
tomatoImage: "certified",
tomatoRating: "6.9",
tomatoReviews: "210",
tomatoFresh: "160",
tomatoRotten: "50",
tomatoConsensus: "With first-rate special effects and compelling storytelling, this adaptation stays faithful to its source material and will please moviegoers of all ages.",
tomatoUserMeter: "61",
tomatoUserRating: "3.1",
tomatoUserReviews: "34104183",
tomatoURL:
"http://www.rottentomatoes.com/m/chronicles_of_narnia_lion_witch_wardrobe/",
DVD: "04 Apr 2006",
BoxOffice: "$291,685,219.00",
Production: "Buena Vista",
Website: "http://www.narnia.com/",
Response: "True"

Metadata:

columns : 28 attributes of a movie.
rows : 5044 movies spanning across 100 years in 66 countries.
There is no spatial data in the dataset.

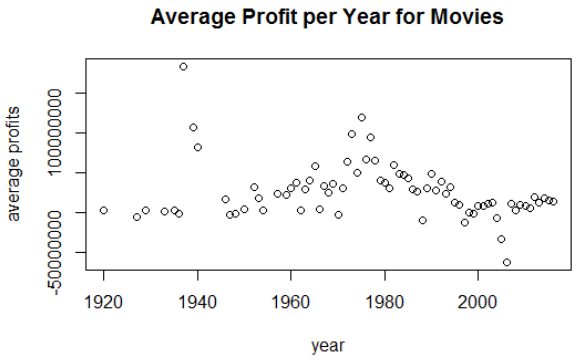
After searching for more movie related data, we came across an API called the OMDAPI that gives consolidated information for a particular movie from different websites. We decided to use this API (not yet implemented) to get data about the reviews from Rotten Tomatoes (a famous movie review website).

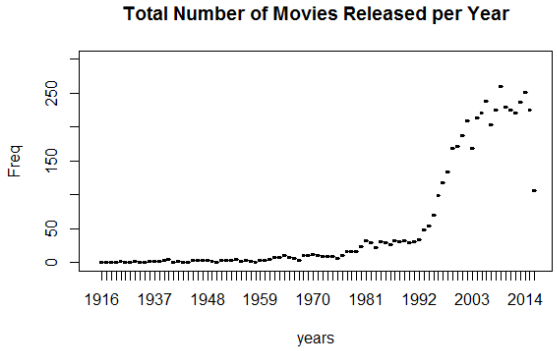
Link: http://www.omdbapi.com/?t=narnia&tomatoes=true.
Sample data provided as a screenshot.

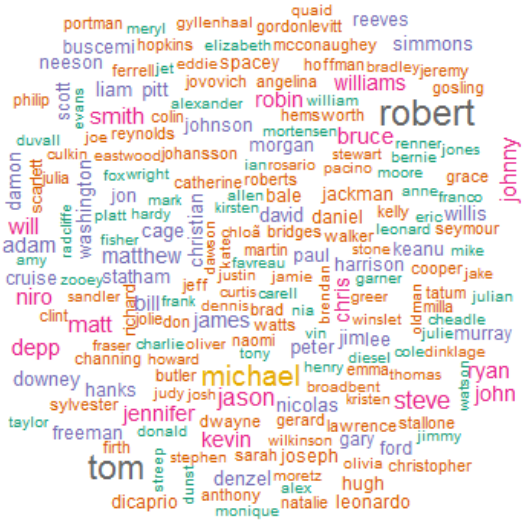
Basic exploratory data analysis:

For better understanding the data, we were looking at different aspects and features of the data set and made some valuable observations that further motivated us towards the project idea.

```
> cor(movie$imdb_score, movie[,c(3,4,29,30,13, 19, 24)], use="pairwise.complete.obs")
     num_critic_for_reviews  duration total_facebook_likes    profits num_voted_users
[1,]              0.3053029 0.2616615            0.2439066 0.03663351       0.4109652
     num_user_for_reviews title_year
[1,]            0.2924754  -0.209167
```







Factors that we consider to rate the success of a particular actor:
(for an individual movie)

1. The budget of the movie
2. The profits made the movie
3. The IMDB rating
4. The Rotten Tomatoes rating
5. The total number of facebook likes.

To summarise all the above factors, we wanted to come with a formula that contains the means of all the above mentioned factors for a particular user, we are yet to experiment of the formula but a naive formula would be : **(mean(profits/ budget) * max(likes in facebook) * max(rating in Rotten Tomatoes) * mean (rating in IMDB)) / total number of movies made by the actor**

Data cleaning:

1. The data is little inconsistent in the column that contains the number of facebook likes, few very famous directors have 0 likes, we wanted to manually fill in these values because we feel that facebook likes contribute 20% to the IMDB score
2. There are a few NA valued cells in the gross column, we decided to remove such rows because none of the missing data approaches felt relevant as the gross is heavily movie dependent
3. Not data cleaning really, but we had to reorder the columns so that one can extract maximum understanding from the data
4. With the use of OMDB API we are planning to add in 3 more columns into the data set namely the “Tomato User reviews”, “Tomato Rating”, “Tomato user rating”.

Outcomes:

For the first problem,

We will be presenting the weights for the strengths/success measure for a particular actor in 10 genres. (We are working on identifying the top 10 genres to choose for the analysis; idea is to extract genres in which most movies are made in the past 60 years). The output variables would be 10 success measures(one in each genre) and a top pick genre for the next movie for an actor.

This outcome is useful because the actor or the director can pick the correct genre that increases the profitability of the movie and the likeliness of the movie.

For evaluating the output, we pick the top genre by picking the one with the highest weightage and compare with the profits and review counts (IMDB and Rotten Tomatoes) of that actor with the movies from the test data set and show that the genre that we suggested was the most successful.

For the second problem,

We will presenting the predicted IMDB score for a movie. The output variable would a single numerical value (range 0 - 10) for a movie which would be its IMDB rating.

This outcome is useful because the users can be given heads up with how the movie is going to fade and manage their expectations for the movie. Also this can be a benchmark for the movie makers to predict how their movie is going to perform.

For evaluation, we will measure the RSS and the R^2 errors between the predicted value and the actual data from the test data and tune the model accordingly by selecting another set of variables

Describing the methods:

1. **Using multi class classification algorithms:**
   We plan to use **Linear Discriminant Analysis**, and using **multi SVM classifiers** along with the **voting mechanism** , etc., Also we want to compare the performance of the above mentioned classification algorithms and the pick the best one for this data set.

For specifics of the implementation, we will be using R packages that does the work, i.e l**da {MASS}** (coming from the MASS library) and **predict.svm {e1071}** (coming from the e1071 library).

We feel using LDA is appropriate because we can ask LDA to give **probabilities** for the data point to fall into a particular class label. We want to compare the success of the actor in all the genres so probability is the key instead of binary classification. Also we read that LDA does a better classification job when compared to multiple binary classifiers in most of the problems.

2. **Using regression techniques for prediction:**

We plan to use **LDA** for prediction (along with the **predict()** function) also along with the use of standard regression models like **lm()** (fitting to be done by the least squares approach).

Assumptions:

i. We are relying on the fact that facebook like, gross revenue and number of user reviews are the best ways to predict the success/rating of a movie and an actor. Although these three variables contribute majorly, there are still other variables that aid in decision making but we are majorly choosing these.
ii. Because the data is not collected by us, we rely on the accuracy of the data collected by someone online. We wanted to manually enter the number of facebook likes for few of well - known directors and actors. So we assume that we are doing a good job here.
iii. Also because we don't have data about the number of positive reviews and negative reviews, we assume that more reviews for a movie correspond to say that the movie is a hit movie
iv. Because the data has data points from 1960 - 2016, it is difficult to take into the consideration the inflation in prices, so we decided to take the ratio of profits to the budget to shadow this issue.