Movie Rating Predictor Final Report

1. Motivation:
   Ever since the first motion picture camera was invented 120 years ago, film, one of the richest form of human expression and a now gigantic industry that generates tens of billions of dollars every year, has witnessed numerous technological and artistic developments and engaged almost everybody with its entertainment. Nowadays, while many people are working towards producing movies of the highest quality, the stakeholders of this industry are increasingly focused on commercial success. As a result, very few studies have been done on how to make high quality films as opposed to how to achieve box office success. Aspired to discover the secret formula of producing the high quality movies, we as a team are going to explore multiple features of film production and advance people's understanding of this topic with machine learning models. We believe our endeavors will benefit many people especially film producers, directors, companies or students who are seeking the magic formula to craft high quality films.

2. Dataset
   IMDB database and Wikipedia serve as our main data sources. We queried the IMDB database for all qualified movies by the following filters:

a. Year > 1976
b. Country = USA
c. Number of Rating Votes > 500

As a result we got a total of 8942 movies with ratings ranging from 1.1 to 9.5. We further filtered out examples that have many missing data (movies that don't have box office, budget, and actor data) to improve training performance and are left with 4476 instances.

IMDb rating is widely used by critics and consumers as an indicator of movie quality relatively independent of commercial success. Therefore we rely directly on IMDB ratings to generate our movie quality label. We used two ways to label/classify instances according to their ratings 1) divide ratings into four quartiles representing four levels,

a. A: rating >= 7.1, first quartile
b. B: rating >= 6.4 second quartile
c. C: rating >= 5.7 third quartile
d. D: rating < 5.7 fourth quartile

and 2) dividing ratings into two groups, below and above top 40% line.

a. A: rating >= 6.7, top 40%, high quality
b. B: rating < 6.7, lower 60, low quality

For each movie, we collected the following attributes:

a. Year (numerical)
b. Main genre (nominal)
c. Number of genres (numerical)
d. Length (numerical)
e. Number of languages (numerical)
f. Box office (numerical)
g. Budget (numerical)

h. Award Index of top three actors/actresses (numerical)

i. Award Index of director (numerical)

Among these attributes, a) to e) can be obtained directly from the IMDB database. For attributes f) to i), we wrote a python script to collect them from Wikipedia. Most of these attributes listed above are straightforward, probably except the last two. The original intention of the last two attributes is to take the number of awards owned by directors/casts into account. However, considering the difficulty of counting the exact number of awards, we decided to approximate it by the frequencies of some award related keywords in the Wikipedia page of directors/casts. Keywords considered for this purpose includes “oscar”, “golden globe”, “prize”, and “awards”. Although such an approximation sacrificed accuracy to some extent, the compromise did bring us the practicability to take this attribute into consideration.

3. Methods (training testing)
4. Results
5. Future Steps

After our experiments, there are a couple of future steps that we would consider:

1. Include more features that have an empirical impact on movie quality including special effect techniques, sound/videography effects, etc. Especially, as movies are usually made by many people with myriads of roles, finding a way to factor all of the human part of movie productions should be very effective.
2. We have seen that award indices of directors and actors have proven to be very effective. We intend to refine our way of calculating the award index of actors and directors. Right now we do a simple count of a few keywords, which may be a good representation of the award concept but does differentiate or give weight to different awards and nomination versus winning status as well as the date of award.
3. Considering that a large portion of movies have missing data on multiple attributes, we may want to expand our data sources so that we can collect more comprehensive data for each movie. By doing so, we can avoid losing around half of the movies simply because they have too many missing data.
4. Experiment with more training methods. Neural network has proven to be the most effective. However, there are still many different settings of this model that should be test with in hope to achieve better testing results.

Actual Report:

Synopsis:

Our task is to predict the ratings of movies from the attributes provided by the imdb database and wikipedia. Imdb database provides well recognized quality indicators for films and movies. Our model gathers the past data from the imdb server (for most attributes) and wikipedia (for box office, budgets, and film stars), learns from the past data to build a decision tree, and then makes predictions for new movies. Movie producers or directors can use our model to predict the rating of future movies from the movie attributes and thus find the magic formula to produce high quality films and avoid low quality ones.