Movie Rating Prediction Report

Purpose and dataset

There are thousands of movies were produced each year in the world. Some of them are great, some of them are bad. We can leverage some useful movie information to predict a new movie rating to avoid watching bad ones. This project is going to predict movies rating and reveal the most significant movie information which influences a movie rating by utilizing machine learning techniques.

The movie_metadata dataset contains 28 variables and 5043 movie records. Half of the variables are directly related to movies themselves, such as title, year, duration, country, genres, language, budget, and gross. The other half is related to the cast which involved in movies and the popularity of movies, such as director name, director Facebook likes, actor name, actor Facebook likes, movie Facebook likes, IMDB scores and so on.

Exploratory data analysis

IMDB_score

Movies rating with more than 8.0 are listed in the IMDB top 250, and they are considered as great movies. Movies with a rating from 7.0 to 8.0 are still good movies. However, movies with a rating from 1.0 to 5.0 are sometimes bad movies.

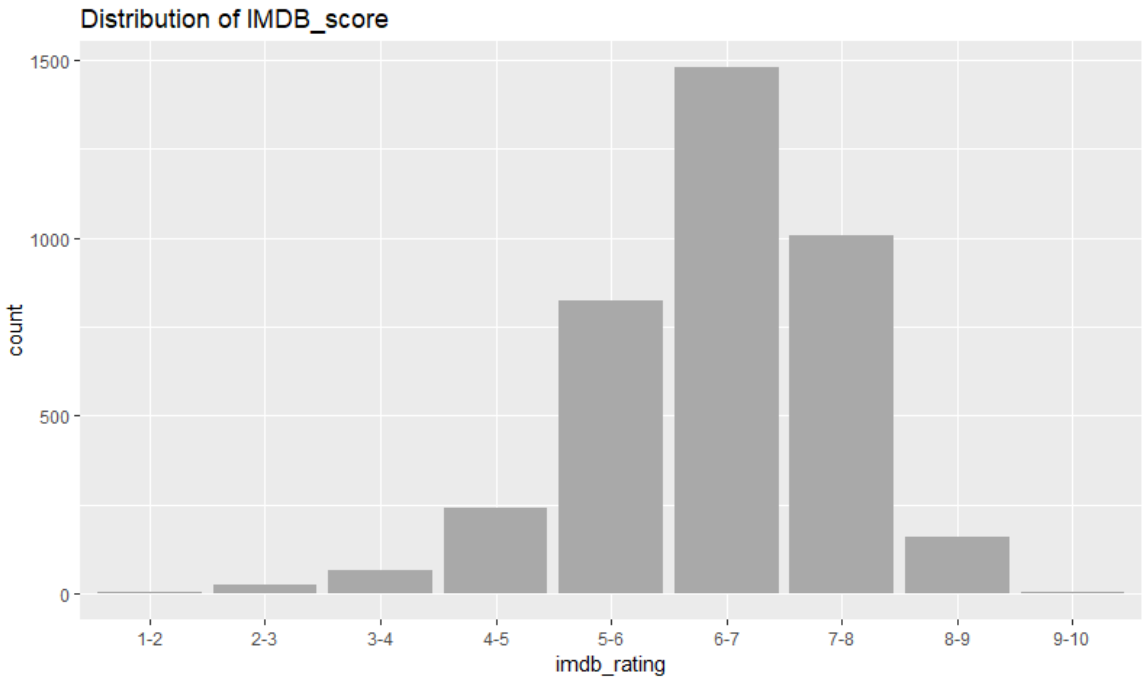

Figure 1

As can be seen in Figure 1, there are a few great movies (8.0-10) but more good movies (7.0-8.0). The majority of movies are scored from 6.0-7.0.

IMDB_rating vs color

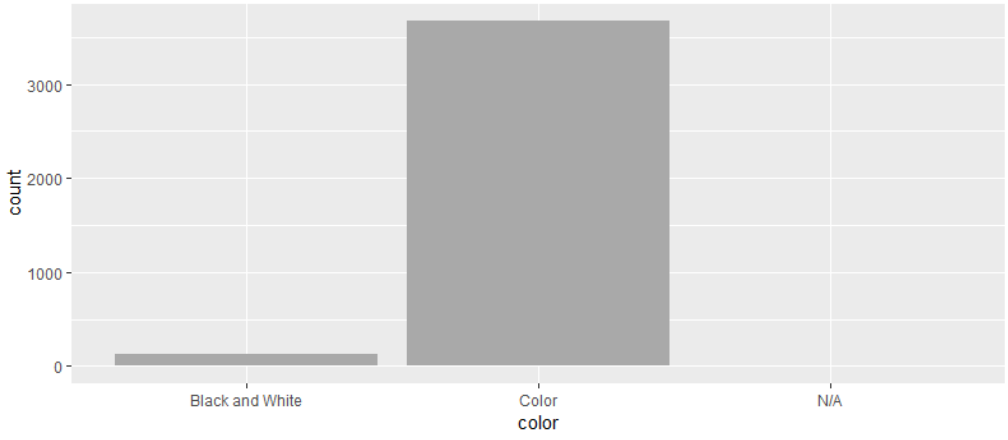

Figure 2 Distribution of color

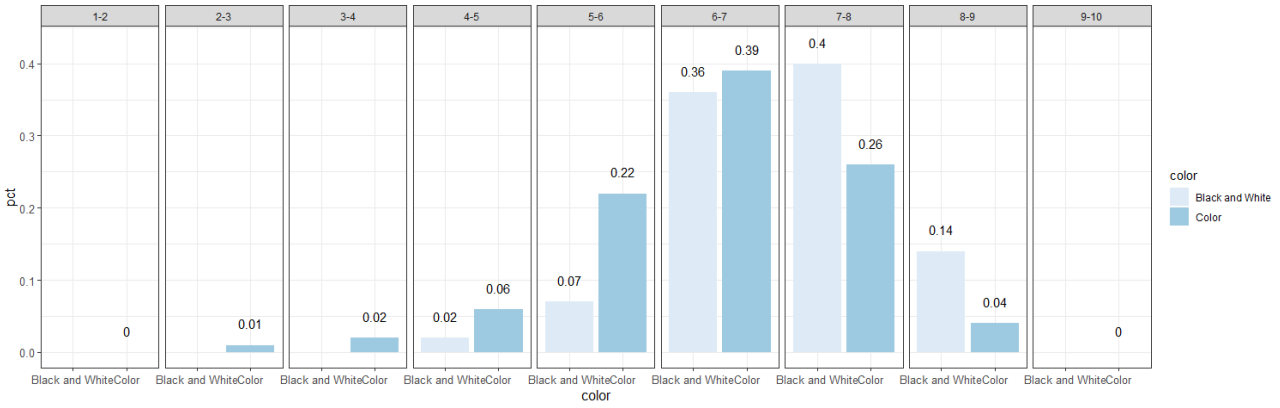

Figure 3 IMDB_rating vs color

Even though there are only a few black and white movies (Figure 2), they are generally good and great movies (Figure 3). 14% of black and white are scored from 8.0-9.0 and 40% of black and white are scored from 7.0-8.0. However, there are only 4% of color are scored as great movies, and 26% of color are scored as good movies.

IMDB_rating vs genres

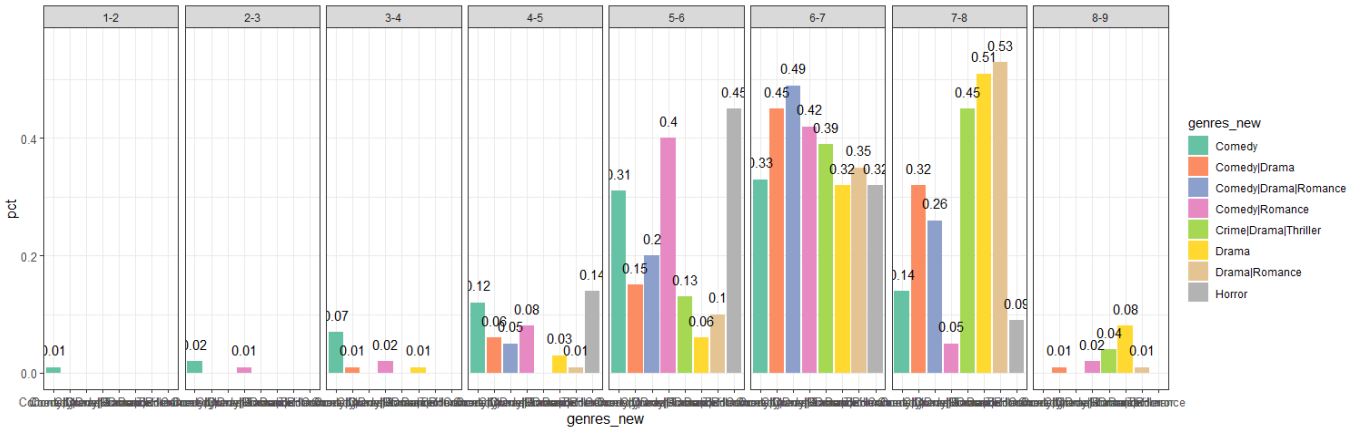

Figure 4 IMDB_rating vs genres

I picked 8 the most frequent movie genres out of 914 genres. Figure 4 shows that 59% of Drama and 54% of Drama|Romance have high IMDB scores. The third is Crime|Drama|Thriller (49%). Certain Comedy movies (22%) sometimes are considered as bad movies.

IMDB_rating vs country

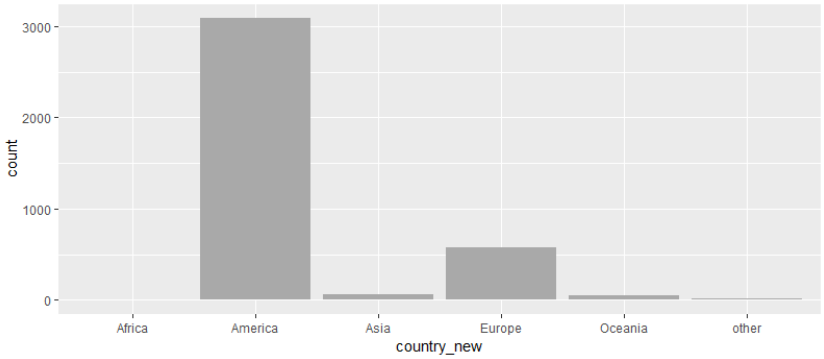

Figure 5 Distribution of continents

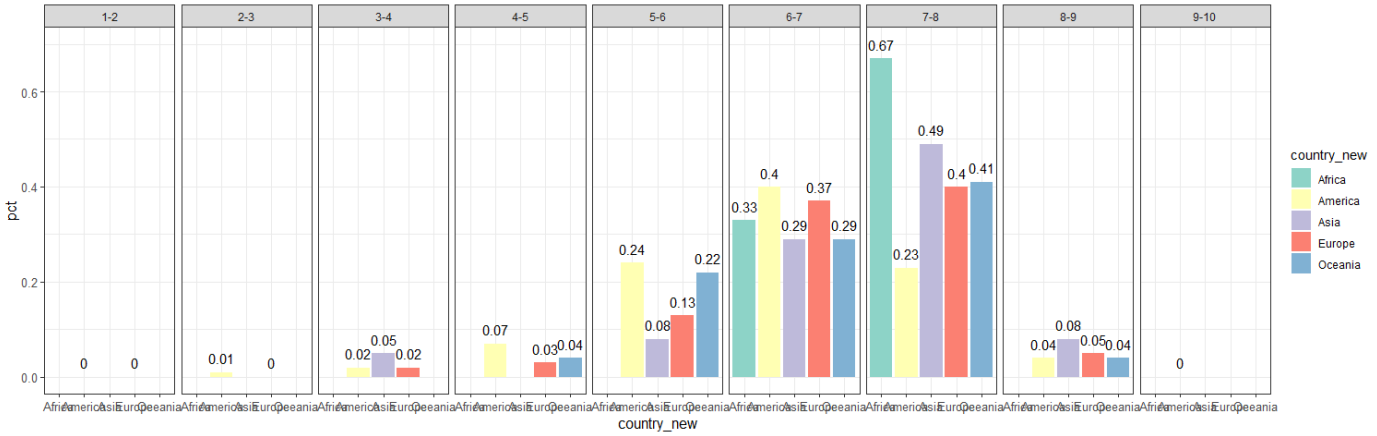

Figure 6 IMDB_rating vs continents

These movies are from 66 countries. I re-arranged them by five continents. The USA and UK are the two countries that produced the greatest number of movies. Therefore, America and Europe are the two continents that contain the greatest number of movies (Figure 5). However, they also have some bad movies (15%) (Figure 6). Africa doesn't produce too many movies, but 67% of them are good movies. As well as Asia (57%) and Oceania (45%). Surprisingly, America produces a large amount of movie, but only 27% of them are considered as good ones.

IMDB_score vs actor_facebook_like

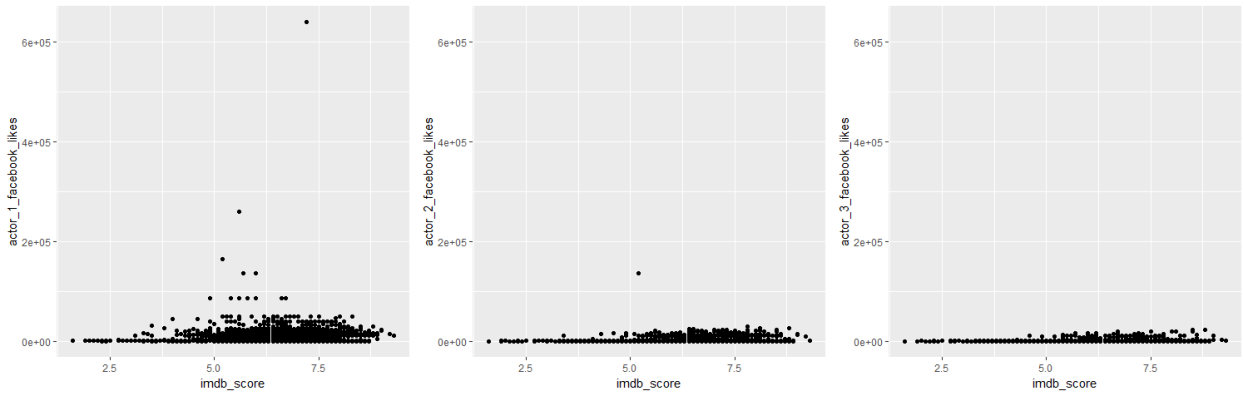

Figure 7 IMDB_score vs actor_facebook_likes

I would like to explore that if the movie would have higher rating if actors have the highest popularity on Facebook. As can be seen in Figure 7, there is no obvious relationship between IMDB scores and actors' popularity.

IMDB_score vs gross & budget

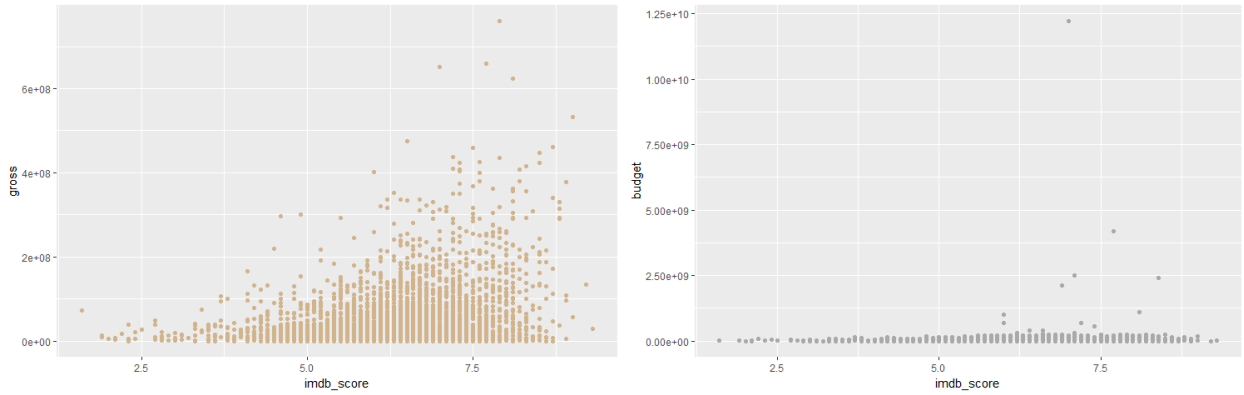

Figure 8 IMDB_score vs gross & budget

From Figure 8, we can see that movies tend to have a high rating if a movie has higher gross. However, budget seems has no influence on movies rating.

IMDB_score vs director & movie_facebook_likes

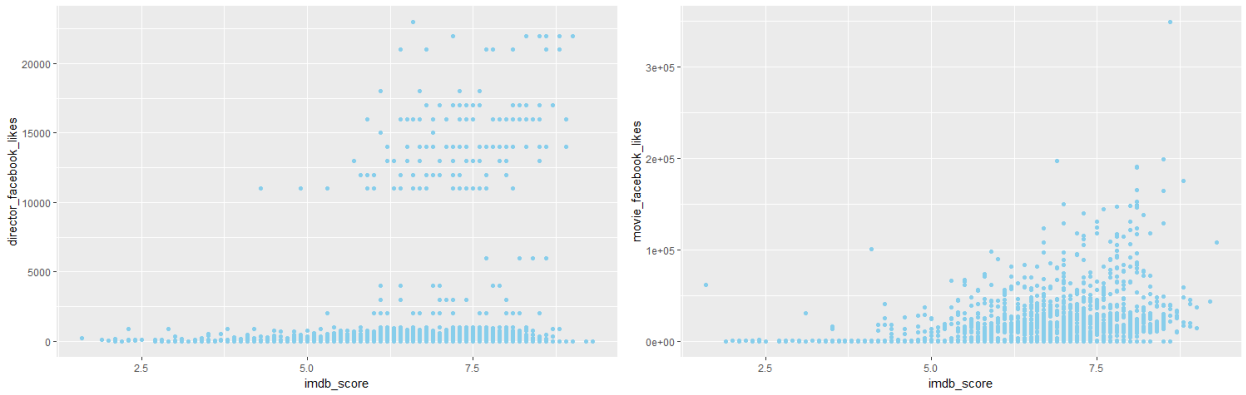

Figure 9 IMDB_score vs director & movie_facebook_likes

We can't tell that a director with higher popularity must direct a movie with a higher rating, because some directors who are not popular but still direct great movies (Figure 9 left). Movies' popularity on Facebook as the same, some great movies don't have too many likes on Facebook, but some do.

IMDB_score vs facenumber_in_poster & title_year

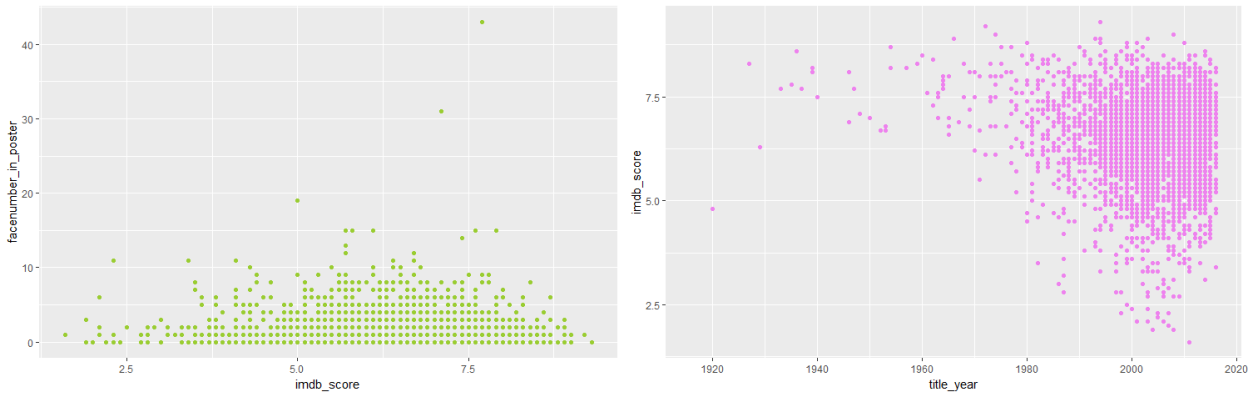

Figure 10 IMDB_score vs facenumber_in_poster & title_year

As can be seen in Figure 10, face number in the movie poster doesn't correlate with movies rating. In addition, as the number of movies increasing largely year by year, some movies with low rating increases as well.

Correlation matrix

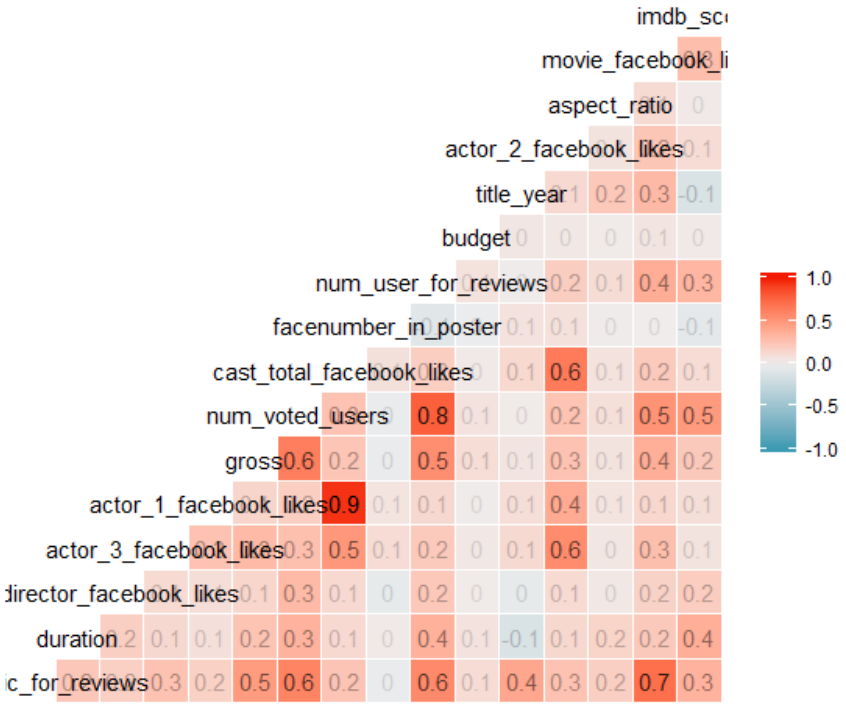

Figure 11

Figure 11 displays which variables have the most significant correlation relationship with movies rating. As we discovered above, title_year and facenumber_in_poster have weak but negative correlation relationship with imdb_score. Budget nearly doesn't correlate with imdb_socre at all. Num_voted_users, duration, movie_facebook_likes, num_user_for_reviews, gross, and director_facebool_likes have weak but positive correlation relationship with imdb_score.

Prediction

I built a training set (75% of samples) and a testing set (25% of samples) for models. The predictive models were used here are Random Forest, SVM, and Decision Tree. The variables

are used to fit models are duration, director_facebook_likes, actor_3_facebook_likes, actor_1_facebook_likes, facenumber_in_poster, budget, actor_2_facebook_likes.

Random Forest

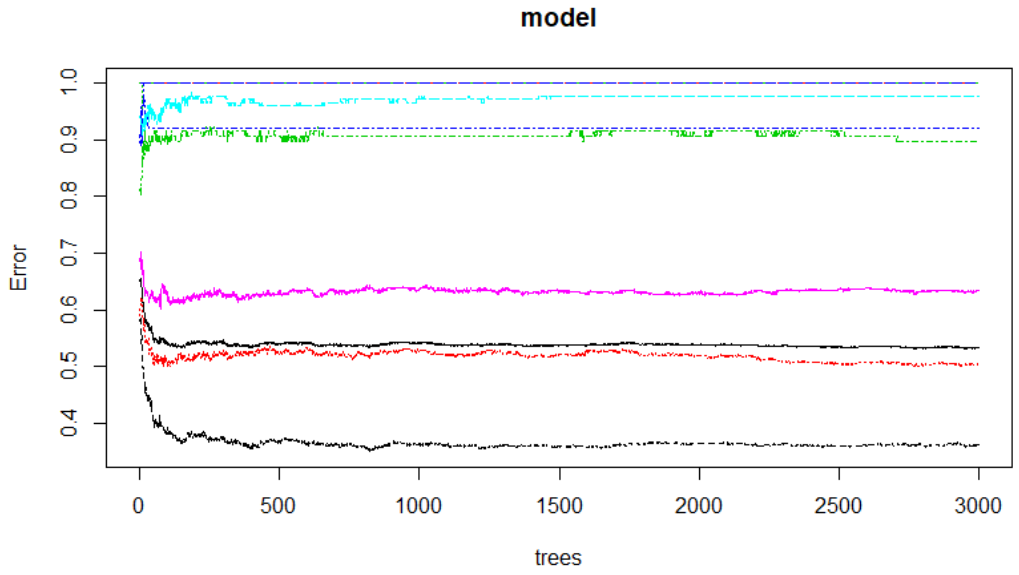

Figure 12

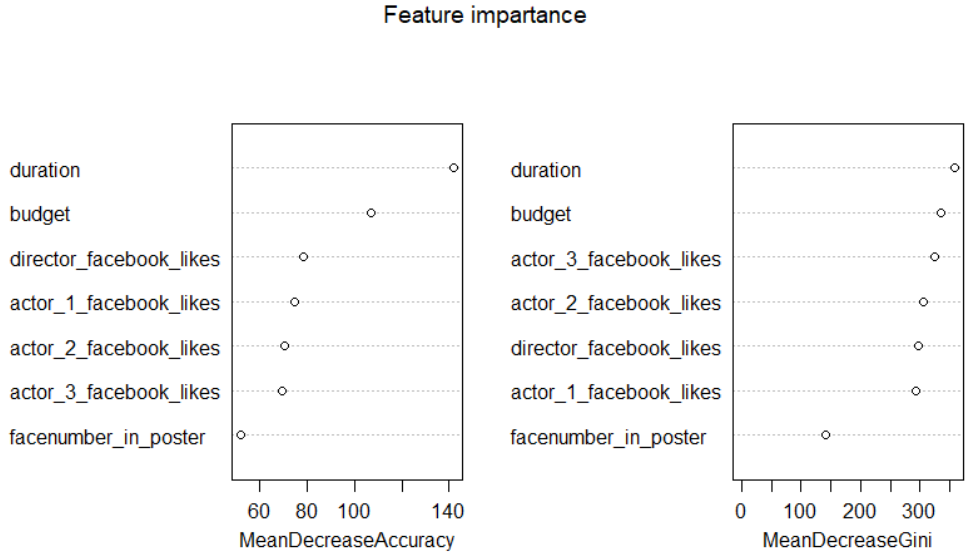

Figure 13

Decision Tree

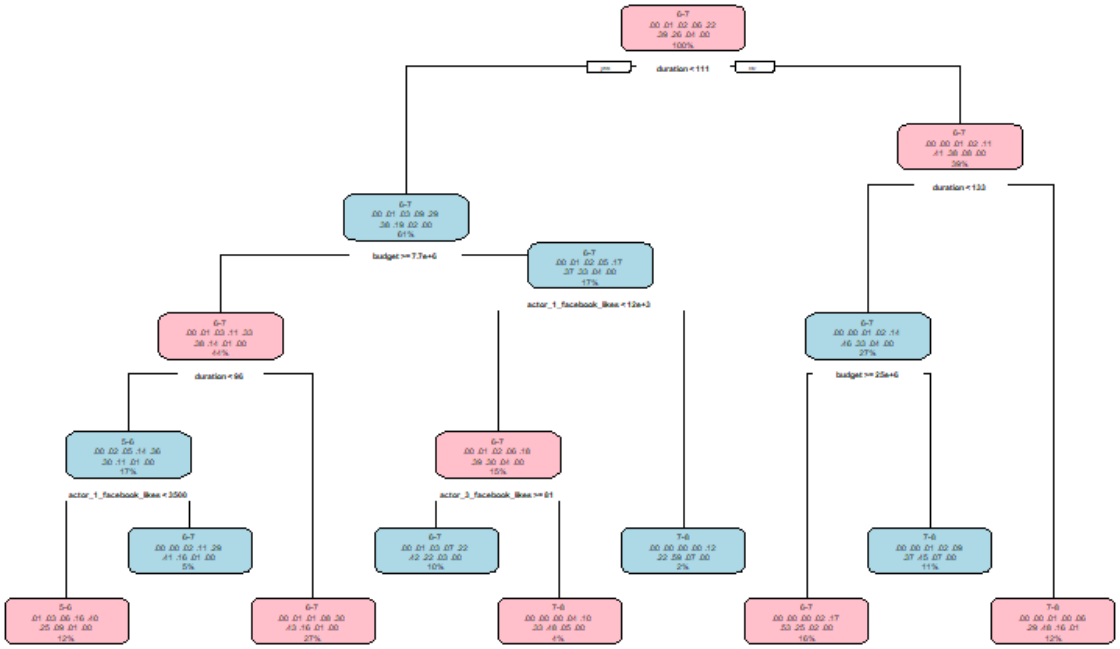

Figure 14

```
> tree$variable.importance
                duration                   budget  actor_1_facebook_likes
               78.613291                44.018361                22.206670
 actor_2_facebook_likes  actor_3_facebook_likes director_facebook_likes
               16.288711                14.752488                 6.186471
```

Figure 15

Findings

- To predict movies rating, we would expect more accuracy of prediction. Therefore, I would suggest the Random Forest model because it has the highest accuracy.
- From the results from Random Forest and Decision Tree, the most important factor that affects movies rating is duration, followed by budget.