Quantitative Method for Market Analysis- R Project - MIM 2016/2017

*Group Members: 1. Nguyen Manh **Luu** (187486) 2. Vo Thi Ha **Giang** (187490) 3. **Michael** Seyram Morkly (187234)*

EXECUTIVE SUMMARY

When asked about what accounts for a good movie, most audiences would think for a rather long time about the answer. This is because there are many factors that determines the "goodness" of a movie, so choosing the best of them would be a complicated task. To some audiences, a good movie is one that is literally impossible to look away from the screen, because every little detail is important and helps to engage the audience deeper into the shown situation. Still to others, it could be a culmination of the budget, director, actor/actress, editing, costume design, set design, cinematography, etc. Whether a movie is a rotten tomato or a brilliant work of art, if people are watching, then it's worth critiquing. Therefore, the major objective of this project was to provide a prediction model based on some variables to enable audiences to determine a good movie (based on assumption that high IMDB score is a good movie) for their entertainment. Life is short!

This project uses movie database of 5000+ movies from IMDB.com, the largest movie database currently to model a prediction tree along with some Exploratory Data Analysis (EDA). The data contains 28 variables for 5043 movies and 4906 posters, spanning across 100 years in 66 countries with 2399 unique director names, and thousands of actors/actresses.

Based on our analysis, we found out that the United States of America was the dominant country with 3805 movies out of the total 5000+. Over the years, the quantity of movies released increased sharply. Steven Spielberg is named as the most popular director. Shawshank Redemption is marked as the best movie per this rating system. Having also considered the most popular actor in the dataset, Robert De Niro emerged on top with roles in 42 different movies. On the other hand, Leonardo DiCaprio was the actor receiving highest average IMDB score for his movies . Surprisingly, Jennifer Lawrence was the only female actress in the top of 20 actors having highest rated movies

Regarding movie budget, since there were differences in currencies, we decided to focus our analysis on the UK and the USA due to the similar currency (USD) in budget. We discovered also that, over the years, the amount spent on movie production increased. This could be attributed to the rise in computer generated imagery and the new millennium which ushered in the age of advanced special effects (3-D and performance capture). We could therefore conclude that, perhaps, most importantly, the movie industry was tremendously influenced by the development in science and technology. Together with the increase in number of movies, there were more great movies with high IMDB score, but also many worse movies were produced. TV-MA (mature content) interestingly is the content rating with highest median scores.

We also used correlation to try to ascertain the relationships between variables. We discovered that, there was a considerable relationship between the number of votes a movie receives from registered IMDB users and the gross revenue of the movie. It can be interpreted that movies that draw more attention from audiences are more likely to succeed commercially. Interestingly, the relationship between investment and profit is rather unclear: some movies with high budget did not do well in box offices. A major drawback in our analysis was our inability to consider currency inflation in general, because 1 USD in 1920 holds a different value for 1 USD in 2016!

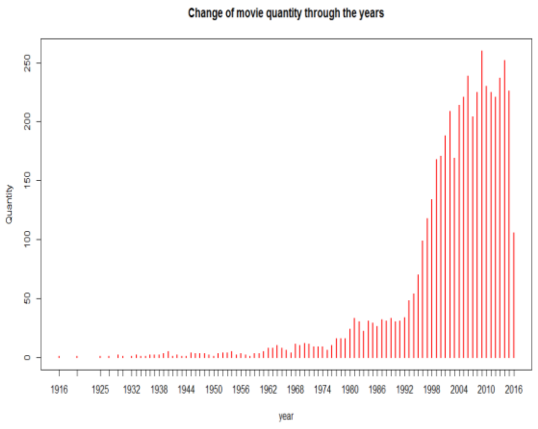
In addition to this, we tried to find a lower dimensional representation of the quantitative variables in the dataset using PCA, but all variables seem to bring some useful information. PC1 places most of its weight on number of critic for review, number of voted user, so PC1 can be a proxy for popularity in general. PC2 is likely to represent IMDB score and budget. PC3 places almost all of it weight on cast total facebook like. In plotting the first 2 principal components, no clear results emerged for group distinctions. We then decided to use h-clustering with different methods of proximity measurement to solve this problem. The results also weren't convincing so we applied the k-cluster instead. This worked quite well where it highlighted a distinct group with all superior indicators (very good movies in this group) but further analysis was needed to improve and find patterns in other groups as well.

Regarding prediction model, we see the improvement from a simple tree to cross validation approach and finally resampling method (bagging and random forest). The result suggests that number of voted users are the most important variable in predicting the quality of a movie, and a movie with more than 560.000 voted on IMDB.com is more likely to be a great movie. Duration is surprisingly a factor in predicting movie quality. In a broader view, public attention (whether good or bad) is highly important for a movie if it wants to be assessed and recognized in terms of quality.

PART1: EDA

1, Country: It can be seen that USA with their leading Hollywood movie industry is dominant in the dataset with 3807 movies compared to cumulative sum 1231 movies of all others countries.

2, Movie number thoughout the years: Generally, during the given period, the number of movie per year saw a significant increase & reached a peak in 2007 with over 250 movies. Data are only collected by June of 2016



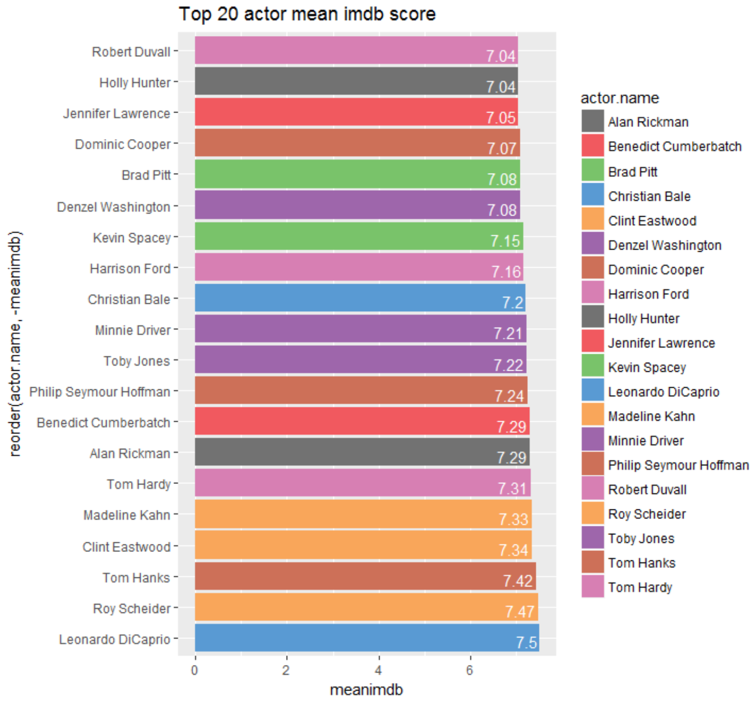
3, Most popular director
Steven Spielberg, as the founding pioneer of Hollywood era, academy winner with various blockbusters like *Jurassic Park, Schindler 's list,*..is the director with the most number of movies (25), followed by *Clint Eastwood* & *Woody Allen* with 19 movies. Two of them come from USA, with the exception for *Woody Allen* (Spain).

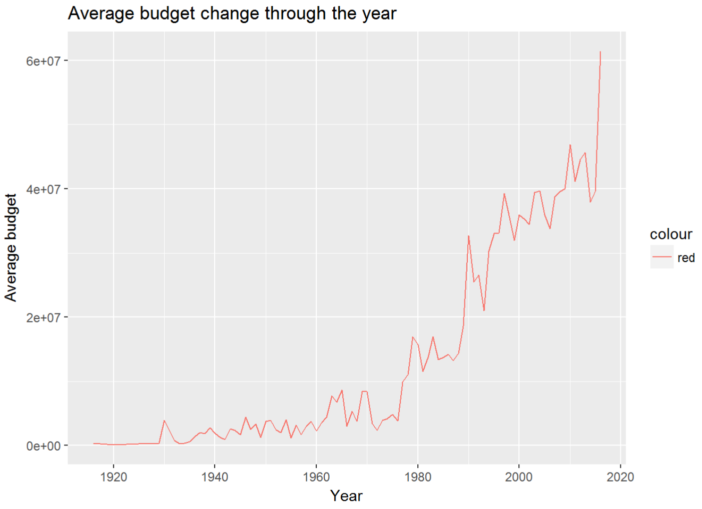
4, Movie with highest & lowest IMDB score: *The Shawshank Redemption* has the highest IMDB score (9.3), whereas the movie *"Justin Bieber: Never Say Never"* only gains 1.6, the lowest IMDB score in the database.

5, Most popular actor, actor and mean IMDB score: *Robert De Niro* was the most popular first actor, playing the lead role in 42 different movies. However, *Leonardo DiCaprio* is the actor having movies that reached highest mean IMDB score of nearly 7.5. Some other famous names also appear like Tom Hanks, Tom Hardy, Benedict Cumberbatch, Alan Rickman. Jennifer Lawrence is a rare female name in the top 20.

6, Movie budget:
From the first run, a Korean movie called *The Host* achieves the top spot, and through a quick search, we see that it is counted in KRW, not USD. Therefore, the following financial will consider only US and UK movies (currency mainly in USD).

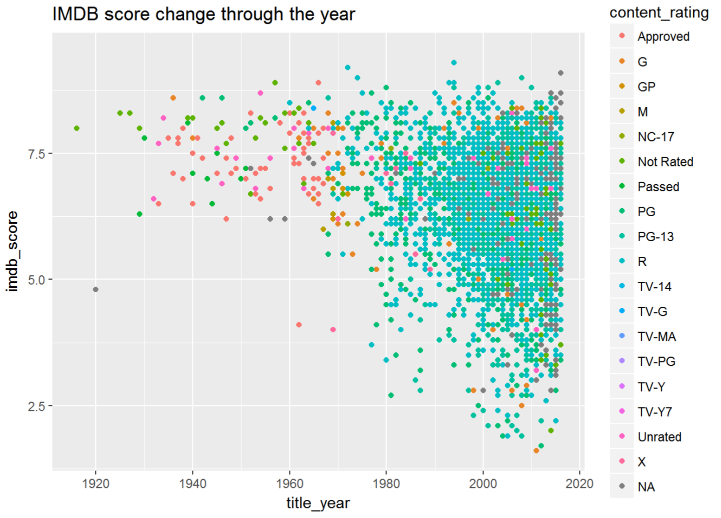
There is a huge gap between top and bottom lines, where highest budget can range from 250-300mil$*(Pirates of the Caribbean: At World's End, John Carter, Tangled* (an animation!)*, Spider-Man 3, The Dark Knight Rises).* Meanwhile, top 5 movies with lowest budget were *Tarnation* ($218), *My date with Drew* ($1100), *Primer* ($7000), *El Mariachi* ($7000), &*Pink Flamingos* ($10000). Surprisingly, these top 5 lowest budget movie received IMDB score from 6-7. Comparing to mean IMDB score (6.466), median IMDB score (6.6), the performance of these low budget movie is not bad. It can be seen that, for some cases, budget doesn't really affect the quality of movie.





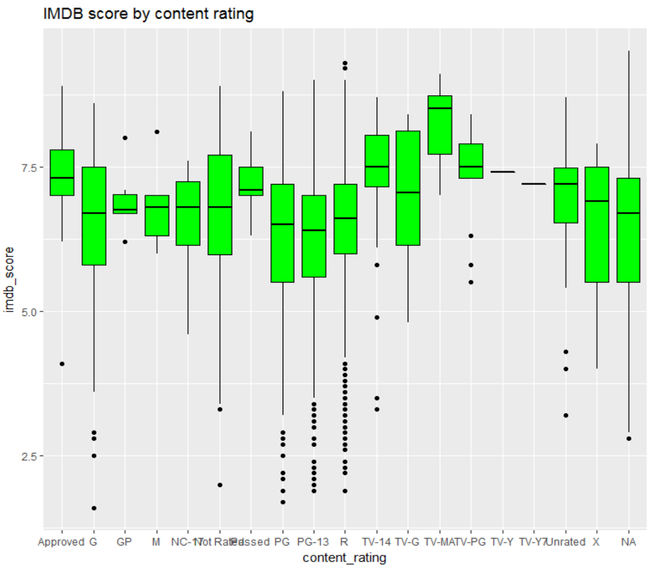
Average budget change through the years: The general trend is increasing throughout the period, but many fluctuations can be seen. Especially, sharp rise can be seen from 1990s, *The Era of Mainstream Films and Alternative or Independent ("Indie") Cinema; and the Rise of Computer-Generated Imagery; also the Decade of Remakes, Re-releases, and More Sequels* and In 2000s, *The New Millennium, an Age of Advanced Special Effects (3-D and Performance Capture), and the Era of Franchise Films.* Obviously these new technology in movies require an increasing amount of investment.

7, IMDB score through years: The number of movies produced annually largely increased from 1970 because the development of filming industry goes hand in hand with the development of science and technology in this period.Through the years,the number of movies gaining higher IMDB scores increased, with more movies concentrated in the range of 6-8 IMDB score. But along with the boom of movie industry since 2000, there are also many movies with low IMDB score. In the period from 1930-1970, there were a lot *Unrated* and *X rated* movies

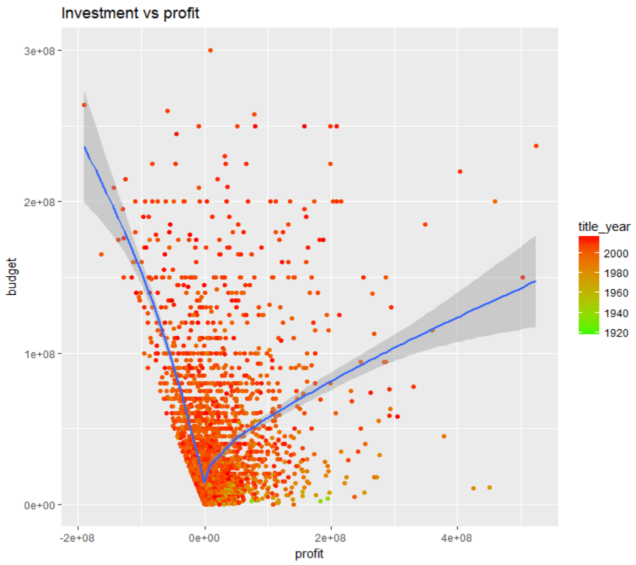


receiving higher IMDB scores.

The box plots with the highest and lowest median belong to movies with content rating *TV-MA* and *PG-13* respectively.Movies with content rating *G, PG, TV-G, PG-13, X* are represented by comparatively tall box plots, which suggests that movies of these categories received the significantly different IMDB scores. By contrast, movies with content rating *GP* , *Passsed* , *TV-14* have comparatively short box-plots, suggesting the rather small differences in IMDB scores among movies of each type. The box plots of *TV-Y &TV-Y7* rated movies are nearly a line, which shows that movies in each group generally gain the same IMDB scores. *PG-13* and *R* rated movies have many outliers , meaning that there were many very bad movie in these categories.



8, Correlation: There is a strong correlation between *num_user_for_review* and *num_voted_users*. This is logical as a user write a review for a movie will also give score rating for that movie. Another strong correlation is *movie_facebook_likes* and *num_critic_for_reviews*. This suggests that the popularity of a movie in social network will attract more critique from critics. *Actor_1 facebook_likes* is highly correlated with *cast_total_facebook_likes,* which is obvious because it is a part of *cast_total_facebook_likes.* A considerable correlation also can be seen between *a movie gross* and *num_voted_users*. It can be interpreted that the number of voted by user is a proxy for the level of audience attention, and the more popularity it gets, the more audiences come to see the movies (increase in gross revenue). Remarkably, the correlation between *IMDB_score* and *director_ facebook_likes/ Actor_1 facebook_likes* is very low, which means that a popular director or actor doesn't necessarily produce a great movie.



Higher investment higher profit?

We have gross profit= gross- budget. Some movies received a negative value for gross profit in spite of very high investment. Most of the movie loss are recent years' movies (in the 2000s onward).

II, PRINCIPAL COMPONENT ANALYSIS AND CLUSTER ANALYSIS.

1, Principal component analysis

We tried to find a lower dimensional represenation of quantitative variables in the dataset. We chose 7 quantitative variables in the dataset, with 2 variables representing the social popularity of the movie(cast total facebook like and movie facebook like) to perform PCA

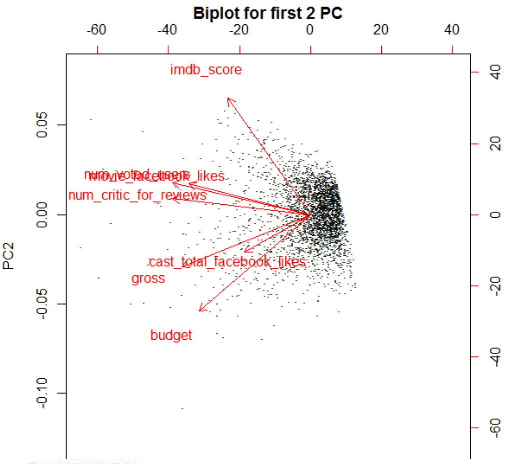
As predicted, PC1 places most of its weight on number of critic for review, number of voted user, so PC1 can be a proxy for popularity in general. PC2 is likely to represent IMDB score. PC3 places almost all of it weight on cast total facebook like.

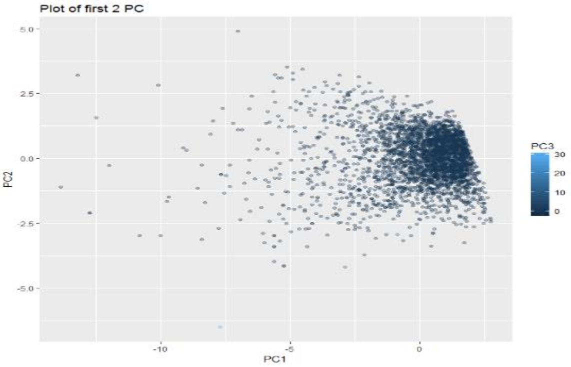
Some outstanding movies in the plot are investigated. Movie *Shawshank redemption* (number 1650) locates separately and in the direction of IMDB score arrow with high score on PC2, which indicates it has a high IMDB score (proved by the fact). *Interstellar* (number 90) is also another case with high PC2 score and high IMDB score

We also tried to define the Variance explained by these 7 PC, by computing the PVE and chart them. The dimensions however cannot be reduced drammatically , because at least the 5 first PCs are needed to explain more than 90% of the data. Therefore it can be concluded that all variables are quite important and bring different information.

We then plotted the first 2 PCs to see if there was any group distinction, but no clear result can be seen

2, Clustering: We tried to define how may clusters was the best option, but the result showed no positive indication, because the *totwithin* does not decrease sharply at an "elbow" when we increase





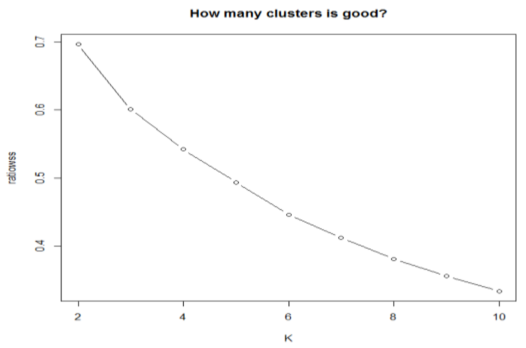
the number of clusters. Then we used h-clustering with different methods of proximity to see how many clusters should be used.

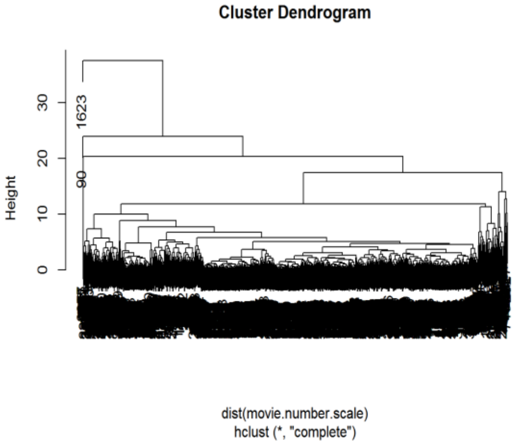
We followed the complete method, we chose 5 separate clusters. We try to apply 5 cluster for K-means. The result of h-clust and k-clust were very different.

H-cluster: There is 1 very big group and other very small groups (with just 1 or some observations only). There may be no clear grouping criteria available using this method or there is no cluster at all from the point of view of h-clustering. However, the observation 1623 and 90 are something different. They are *Interstellar (IMDB 8.6)* and *Anchorman: The Legend of Ron Burgundy (7,2)*. Interestingly, *Interstellar* is the movie with highest *number of facebook like* and *The Legend of Ron Burgundy* has the highest *cast total facebook like* in the whole dataset. This may be the criteria that h-cluster algorithm use to distinguish these 2 movies from the big group. However, overall, the h-cluster here is not so effective in giving a proper clustering solution

K-cluster, the first group have all mean index higher than the other group, so we did further analysis for this group. Obviously, very famous movie with high ranking, budget and gross appear in this group like *Avatar, Spectre, The Dark Knight Rises, The Godfather, Skyfall, Whiplash.* This can be a group of excellent movie with extremely high ranking in both IMDB score and budget as well as revenue.

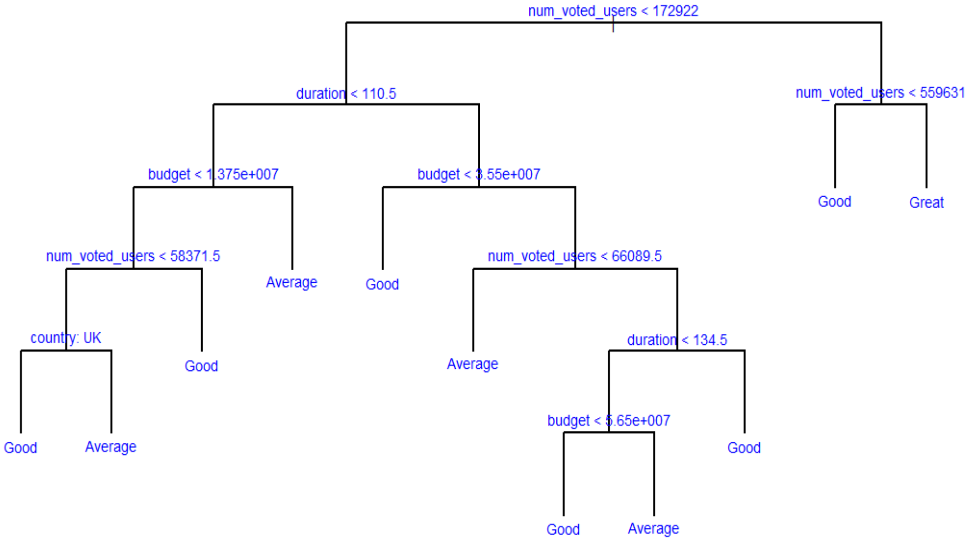
K-cluster works better where distinct group can be found, but further analysis and improvement can be done to find the pattern in other groups as well. Also the *between_SS/total_SS* = 50.6 % is not high.





III, TREE-BASED PREDICTION:

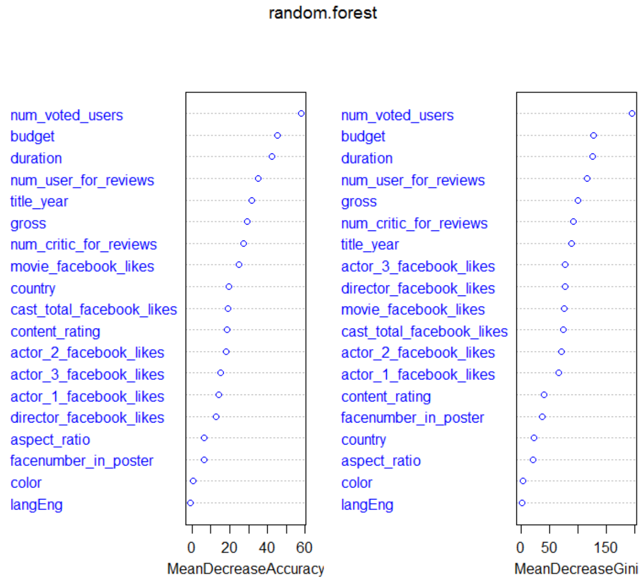
We built a classification tree for IMDB score. We changed the IMDB score from a regression to a classification problem because it was more general for movie audiences to give comments on a movie rather than mention its score. Therefore, movie will be classified based on their IMDB score, based on assumption that the higher score, the better the movie, into 4 classes: *Great* (score>8 - All movie I the top 250IMDB of all-time all get score higher than 8), *Good* (6,5-8), *Average* (5-6.5) and *Bad* (<5). Some variables are omitted either because they does not provide input for prediction or their levels are too many which cannot be executed in R. We followed the validation approach, where we divided the data into 1000 for test set and balance for training set.



The first tree is built with 9 terminal nodes and misclassification rate is 35.7%, test error is 37.3%. The *number of voted users* obviously the most important variable in predicting quality of movie. Surprisingly, *movie duration* is also play some role in predicting. Another point worth noting is that, very interestingly, in some nodes, the tree give prediction that lower *budget* movie will be a *"Good"* movie, higher budget will be just an *"Average"* one.

Then we try with a bigger tree with 223 nodes and 17 variables are considered for node splitting. The misclassification rate reduces to 15.38% and test error rate is 38%, higher than previous tree. This indicates that it is overfitting the model. Then we do cross validation to find out that the best number of nodes is 144 nodes. This is quite a big tree, but we also find that if we decrease the number of nodes to 11, the increase in error not so considerable compare to tree with 144 nodes, but it is easier for interpreting. Therefore, the best node choice is 11, with test error rate improvement down to 35.6%

We finally try the approach of bagging (m=p=19) and random forest (m=4) to see if there is any improvement. Overall, the test error rate of these 2 methods are improved with 30.1% for bagging and 31% for random forest respectively.


random.forest
num_voted_users
budget
duration
num_user_for_reviews
title_year
gross
num_critic_for_reviews
movie_facebook_likes
country
cast_total_facebook_likes
content_rating
actor_2_facebook_likes
actor_3_facebook_likes
actor_1_facebook_likes
director_facebook_likes
aspect_ratio
facenumber_in_poster
color
langEng
MeanDecreaseAccuracy
num_voted_users
budget
duration
num_user_for_reviews
gross
num_critic_for_reviews
title_year
actor_3_facebook_likes
director_facebook_likes
movie_facebook_likes
cast_total_facebook_likes
actor_2_facebook_likes
actor_1_facebook_likes
content_rating
facenumber_in_poster
country
aspect_ratio
color
langEng
MeanDecreaseGini