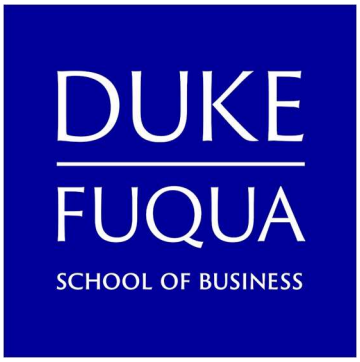
Duke University

DEC520Q Final Project

What Makes a Good Movie?
Prediction of IMDb Scores

DEC520Q - 10B - Team 32



Mutong Zhang
Noah Nguyen
Vainika Choudhary
Xinge Li
Zipei Yang

2019-10-13

I. Introduction

The project's intended audiences are movie producers at the outset of their decision-making process. The dataset used was scrapped in December 2017 from the IMDb website and published on Kaggle.com by Yueming Zhang. Our data compiled information about 5,043 movies with regards to 28 different variables detailing multiple aspects of films (Appendix 1.) We identified IMDb Score as the response variable among other explanatory variables.

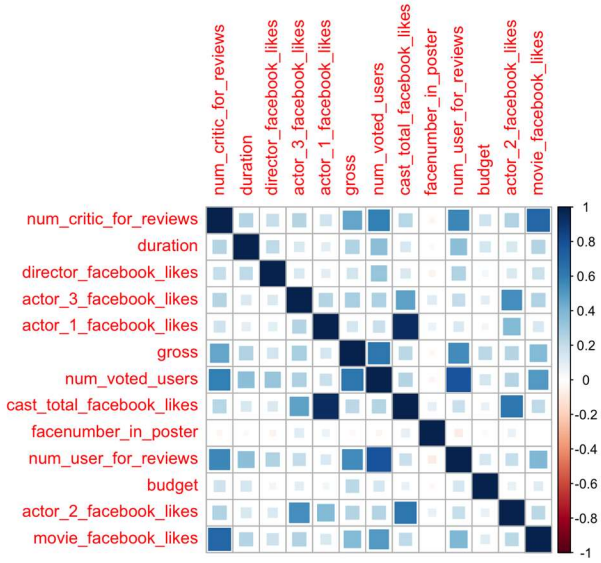
II. Business Understanding

The core objective of our project is to assist movie makers in leveraging existing characteristics of their production to optimize audience reception, presented in the form of IMDb rating score. Though many consider film as a form of art that is difficult to be comprehended and quantified, by building predictive models, we expect this project to discover general patterns in film production that help movie makers understand crucial aptitudes and creative choices that can be translated into well-acclaimed pieces of art among general audiences and film experts alike. Among the variables that are representative of a movie's success, such as Gross Revenue or number of Movie Facebook Likes, IMDb Score was chosen for prediction because it is an indicator that is definitive of audience reception and production quality, and is also emblematic of general commercial success. With input from a platform of 83 million registered users, our final models will give movie producers and investors an idea of which features to consider while predicting movie quality, and how to subsequently allocate their budget and outline marketing strategies (social media, theater selection, franchise, etc.) that are suitable for their movie quality. We will show the results of this analysis in an intuitive way by using various means of visualization, evaluating Regression and Classification models, and presenting effectual implementation guidelines.

III. Data Understanding

Out of the explanatory variables, Movie IMDb Link and Aspect Ratio were excluded from data mining, as we find a movie's link not informative of quality, and that aspect ratios have average impact on IMDb Score (included in R code.) We checked for multicollinearity and noticed that

'actor_1_facebook_likes' and 'cast_total_facebook_likes' almost had the same effect, and 'num_user_for_reviews' and 'num_voted_users' had extremely high correlation. Since Total Cast Facebook Likes can be explained by that of actors 1, 2 & 3, and Number of User Votes pertained more to the explanatory variable than Reviews, we excluded 'cast_total_facebook_likes' and 'num_user_for_reviews' from our models.



Graph 1. Correlation Plot

IV. Data Cleaning

Our first approach to data cleaning was to find duplicate rows within the dataset. After deleting duplicate rows, we still found that certain rows were referring to the same movie but had slightly different values on certain variables. This may be due to the effect of grabbing data of the same movie at different times, and in effect, the numbers recorded are different. In this aspect, we cleaned and left the records referring to the same movies with the largest value of data recorded.

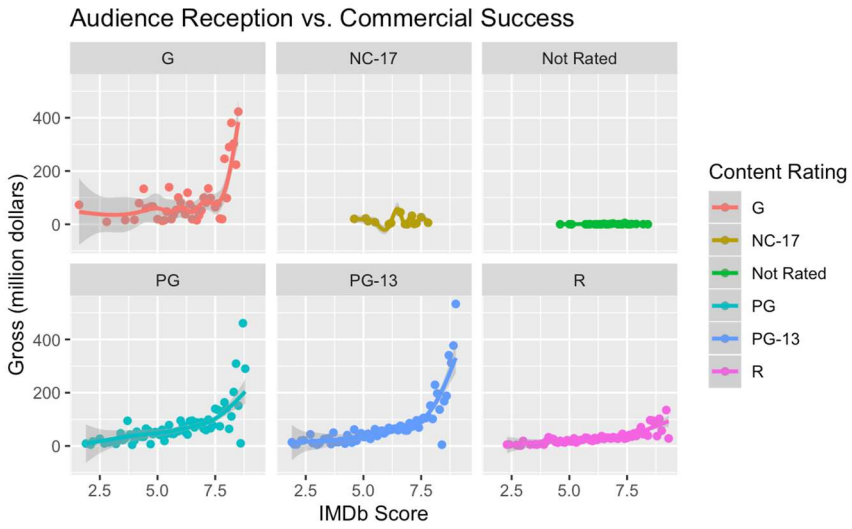
The second step was to find 'NA' values, and we recognized columns like Budget, Gross, Releasing Year (title_year), IMDB Link (movie_imdb_link), and Aspect Ratio had missing values. For Budget, Gross, and Releasing Year, since their values and distributions vary considerably for different movies across time, we believe a biased effect would take place if we simply impute missing values with

summary statistics of other data points. In this aspect, it would be proper for us to take out data points with missing Budget, Gross, and Year of release, in order to get an unbiased dataset. However, about 17% of the observations still had missing Gross values, so instead of obliterating them, we created a sub-dataset and built models for it specifically for further exploration. As mentioned in the previous section, columns like IMDB Link and Aspect Ratio will not be put into consideration for this project, thus they were removed directly. At this point, the column of Number of Faces in Poster still have missing values, but we consider it reasonable to input the mean value of the column back into them. Beside 'NA' values, we also identified values of '0' in columns like Number of Facebook Likes for directors and actors, Duration, Number of User Reviews and Faces in Poster. For these columns, we also replaced '0' values with the mean values of corresponding columns.

For easier further date exploration, we have done text cleaning, like reformatting the storage of strings and removing meaningless punctuations that were systematically created when the data were scrapped, for columns of Title, Genre, and Plot Keywords. For the column of Content Rating, some ratings standards are outdated or based on the TV platform, so we have changed and grouped them into six modern standards of movie content rating (G, NC-17, PG, PG-13, R, and Not Rated), while creating a level of "Not Known" for 'NA' values in this column. Lastly, for the column of Language, since only 4% of the data are non-English movies, we grouped them into the value of "Non-English" for segmentation purpose. To this point, the cleaned data has 3,789 observations with 27 variables.

V. Explanatory Analysis

1. Visualization



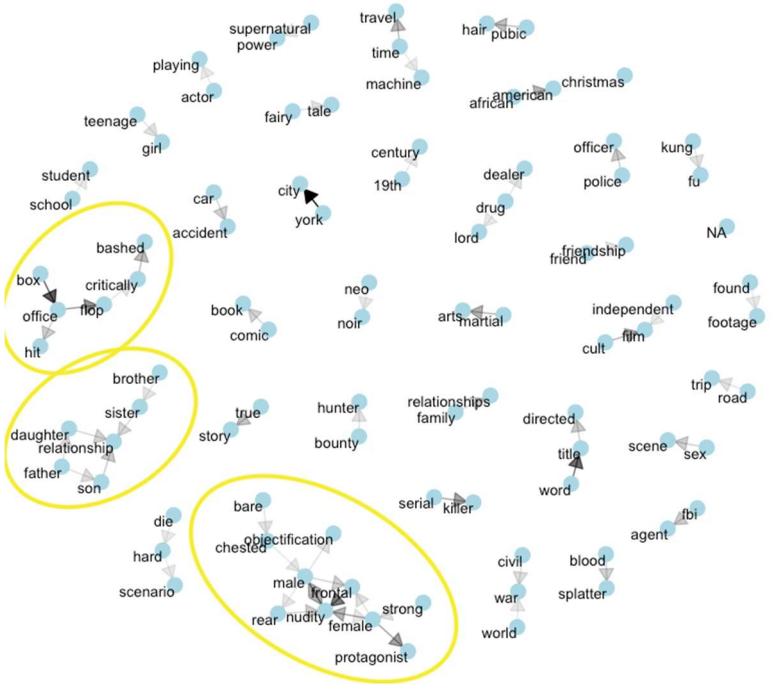
Graph 2. Audience Reception vs. Commercial Success

To further validate our choice of the explanatory variable, we plotted it against Gross - which represents a movie's commercial success - for known film content ratings. Not Rated movies are small-scaled productions that are not meant for theater showings (otherwise they would have had a content rating), so producers of this type should also use IMDb score as a clue for success. We can see that for G, PG, and PG-13 rated movies, there is a strong, upward relationship between Gross and IMDb Score. This relationship is also noticeable in R-rated and NC-17 movies, meaning that for most productions, acclamation is synonymous with box office victory. Investors can rely on improving movie quality to anticipate profitability.

2. Text Mining

We also utilized text mining techniques to visualize a directed network of bi-grams appearing most frequently among Plot Keywords. Based on the weight of arrows, pairs of words that show up most often together are 'york'- 'city' (as many movies center around New York City), 'box'- 'office', 'frontal'- 'nudity', and 'female'- 'protagonist'. These are the trends that other movie marketers want to highlight from their films, and are common among current productions. From dense clusters (circled), we can see that Plot Keywords also feature a production's performance, that many movie plots focus on family relationships, and a lot more talk about male - female relationships. It is up to the production crew to

produce a movie that is either common-themed (easily identified by audience) or niche (novel, fresh to audience perception.)



Graph 3. Network of Plot Keywords Bigrams

VI. Predictive Modeling

1. Regression

The core task we focused on was using regression methods to predict IMDb Score. This is a regression task. According to previous reasoning, we selected ‘color’, ‘num_critic_for_reviews’, ‘duration’, ‘director_facebook_likes’, ‘actor_3_facebook_likes’, ‘actor_1_facebook_likes’, ‘gross’, ‘num_voted_users’, ‘facenumber_in_poster’, ‘language’, ‘content_rating’, ‘budget’, ‘actor_2_facebook_likes’, and ‘movie_facebook_likes’ for modeling.

We used Lasso, Post-Lasso and simple linear regression methods to evaluate performance. Within the Lasso method, we used two models with different values for lambda(ƛ) based on data-driven choices, minimum choice (method 'c') and 1se choice (method 'd'). Theoretical choice of lambda(ƛ) was discarded because it only applied to classification class. After that, we employed the Post-Lasso method, which

refits the model after variable selection by both Lasso methods. The refitted linear model selected by Lasso with lambda(ƛ) of minimum choice is method 'a' and the refitted linear model selected by lambda (ƛ) of 1se is method 'b'. For the linear model, we applied simple full regression with variables mentioned above, denoted by 'e'. We utilized k-fold cross-validation (10-fold specifically) to choose the best model among Lasso, Post-Lasso, as well as Linear Regression methods. (Values for them are provided in the Figure below.) The metric used to measure model performance is out of sample R^2(OOS). Using 10-fold cross validation, we deduced that the model ‘c’ wins the majority of the round with highest R^2 values, thus we can conclude that model ‘c’ fits the overall population the best.

Model ‘c’ is a model based on Lasso with a minimum choice of lambda(ƛ). We removed some non-significant estimators and used our final model to predict the test set (included in R code.) The following graph details information for our model:

```
Call:
lm(formula = My ~ ., data = lasso_train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9380 -0.5023  0.0771  0.6362  2.2133

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               4.814e+00  8.743e-02  55.067  < 2e-16 ***
num_critic_for_reviews    2.461e-03  1.438e-04  17.118  < 2e-16 ***
duration                  1.235e-02  7.768e-04  15.894  < 2e-16 ***
director_facebook_likes   3.717e-05  5.707e-06   6.513  8.6e-11 ***
facenumber_in_poster     -2.301e-02  8.943e-03  -2.573  0.01013 *
`languageNon-English`     6.706e-01  8.034e-02   8.347  < 2e-16 ***
content_ratingG           1.123e-01  1.074e-01   1.045  0.29590
`content_ratingNC-17`     4.524e-02  2.423e-01   0.187  0.85188
`content_ratingNot Rated` 3.096e-01  1.341e-01   2.309  0.02101 *
content_ratingPG         -1.096e-01  5.038e-02  -2.176  0.02964 *
`content_ratingPG-13`    -3.812e-01  3.746e-02 -10.176  < 2e-16 ***
`content_ratingNot Known` 4.253e-01  1.454e-01   2.925  0.00347 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9006 on 2988 degrees of freedom
Multiple R-squared:  0.2643,    Adjusted R-squared:  0.2616
F-statistic: 97.59 on 11 and 2988 DF,  p-value: < 2.2e-16
```

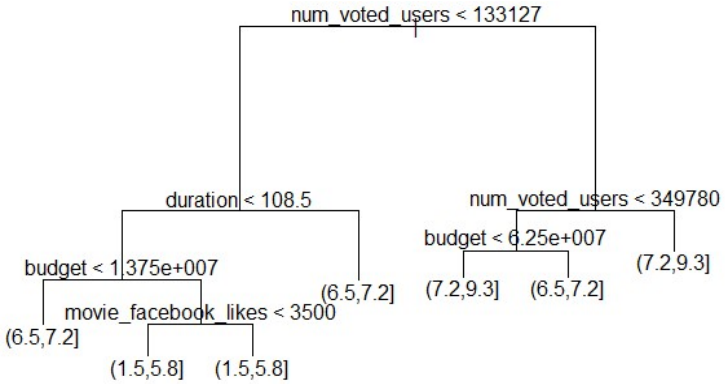
Graph 4. Results of Regression Models

It can be understood from this model that higher-rated movies are those that have more critic reviews, are longer, have more popular directors, have fewer faces in poster, have languages other than English, and are rated for more mature audiences.

2. Classification

Another core task we have identified to be informative is classification, which is essentially

different from regression and can possibly alleviate regression's shortcomings. We used two classification tools: Classification Tree and Random Forest. Before using these tools, we decided to drop categorical variables that have more than 50 classes. These consist of 'director_name', 'actor_2_name', 'actor_1_name', 'genres, title_year', 'movie_title', 'actor_3_name', 'plot_keywords', "language" and 'country'. This was because the CT/RF packages do not allow such variables to be included in the trees, and it makes no sense for producers to use these categorical variables in their prediction. We also used the same train/test split as in regression predictive modeling. Finally, we assigned all ratings into 4 "rating levels" based on quantile information for better clustering purposes. The ranks range between 1.6 and 9.3.



Graph 5. Classification Tree

We then ran Classification Tree first to see if specific features stood out as very predictive. According to the tree above, movies with more votes tend to have higher ratings. For those movies with fewer votes on ratings, a higher-budget and short movie could also earn higher ratings potentially. To further verify significant variables in clustering, we ran random forest on those features and tried to rank them based on importance in predicting which "rating level" a movie would fall into. As stated in the ranking chart, again we confirm the Number of Voters, Duration and Budget remain the three key factors in clustering into the rating level.

```
> rankImportance
                 Variables Importance Rank
1                    color       6.43  #15
2   num_critic_for_reviews     174.71   #7
3                 duration     216.86   #2
4  director_facebook_likes     175.24   #6
5   actor_3_facebook_likes     159.23   #8
6   actor_1_facebook_likes     146.54  #10
7                    gross     198.44   #5
8          num_voted_users     289.20   #1
9     facenumber_in_poster      76.54  #12
10    num_user_for_reviews     200.15   #4
11               language      18.68  #14
12         content_rating      74.29  #13
13                 budget     207.56   #3
14  actor_2_facebook_likes     153.72   #9
15    movie_facebook_likes     144.69  #11
```

Graph 6. Ranking of Variable Importance in Prediction

As discussed before, we tested the subset of rows missing Gross values on Classification Tree and Random Forest as well, and it returns nearly identical results (see Appendix 2.) We then concluded that dropping those rows did not further impact the validation of our current model.

VII. Evaluation

Regression

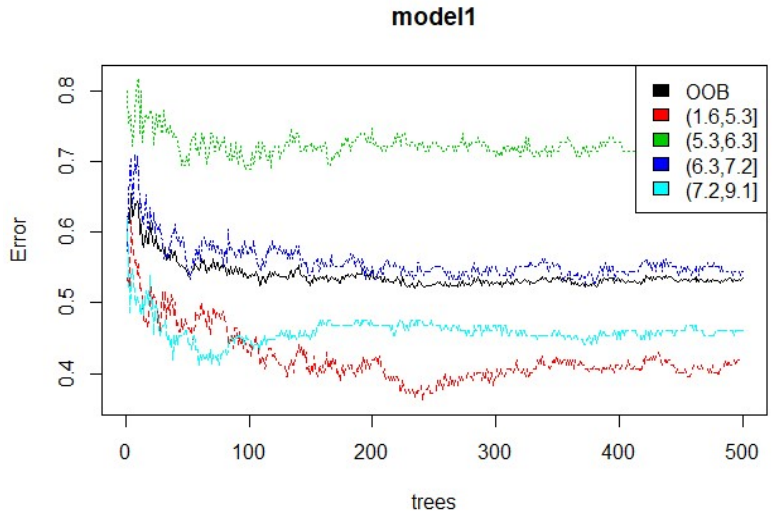
We compared our predicted values from regression with the actual values and calculated the average error for our model, which came to 0.6328603. In other words, the actual IMDb rating for a movie will be in the range of +/- 0.6328603 of our predicted value on average.

We believe this model can provide movie producers with an approximate methodology to predict their movie's success, but a linear regression model with Lasso method does not totally explain the relationship between movie rating and movie attributes, as only 26.43% of the data can be explained by this model. It can be reasonably assumed that there are other non-linear models (such as Classification Tree) that maybe more suitable for our dataset. Another way to improve the accuracy of this regression model is to extract other variables from the IMDb website, such as top box office movies shown in the same period, camera/film reel type, written reviews (utilizing bag of words), etc.

Classification

Neither the Classification Tree nor Random Forest returns optimal results in terms of accuracy. The misclassification error rate for Classification Tree is 56.4%, relatively high compared to other predictive models. The Random Forest model yielded an error rate of 49.4%, a minor improvement from

the Classification Tree but still not optimal. However, we found our model was more accurate when predicting movies in the top and bottom rating levels, but moderate in the middle tiers, as illustrated in the following OOB Error graph.



Graph 7. OOB Errors

To improve the accuracy of the classification models, we believe adding more important features to the model will generally give more accurate predictions. Also, the current clustering criteria (rating level) might not be the best way to group the movies. Without the scope limitation, other success determinants such as ROI might be a more informative clustering mechanism for us to gain insight from.

VIII. Deployment

The results of our regression model suggest that Number of Critic Reviews, Duration, Director's Facebook Likes, Number of Faces in Poster, Language, and Content Rating can influence a movie's IMDb score, while our classification results suggest a Number of Voted Users, Duration, Budget, and Movie Facebook Likes. Out of these factors, some can be decided within the creative process, some within the human resources process, and some within the marketing process.

Within the creative process, producers can decide to make longer movies with bigger budgets, aim for more mature audiences, and invest in non-English productions. Language is a less flexible choice, as it also concerns a capable production crew, possibly different filming locations, and other regulations and tax terms. However, targeting mature audience with higher rated content, reserving more budget, and

investing in more profound screenplays and ideas to ensure longer duration are easier choices to make, though they might conflict with a production company's creative direction. (i.e. Disney will be unlikely to produce R-rated movies.) To mitigate this, production companies can delegate subsidiaries to focus on different creative styles to diversify their repertoire. (i.e. Disney also owned Miramax Films.)

Within the human resources process, it is more beneficial for movie ratings if producers can employ more popular directors. In-demand directors will certainly cost more to employ, so there is a tension between balancing director's pay with other components of the budget. There is also potential clash in creative direction between the production company and the director, and director's availability compared to production timeline. Moreover, colleague circles exist within the creative community, so producers may not be able to secure a director without hiring her trusted scriptwriter or camera operator, etc. This also branches into an ethical concern that is prevalent in Hollywood - production diversity - as purposefully choosing particular directors may result in a monolithic production team. Hiring acclaimed directors will bring media buzz and prestige to a film's production, which have a positive effect on its IMDb Score, but this choice is not always feasible.

During the marketing process, producers can determine to put fewer faces on a movie's poster, and invest more on pre-showing marketing initiatives to ensure higher ratings when the movie comes out in theater. By investing in marketing channels such as a movie's Facebook page, producers ensure that more people know about the movie and show up in theaters, and subsequently vote for it on IMDb. With more pre-show buzz from effective social media marketing, a movie is more prone to a higher rating.

Beside decision-making value for producers, our models present patterns of successful movies for investors and sponsors (product placement) to identify a quality movie. The two parties should take into consideration these factors to optimize the brand value of their investment, and invest in a good movie.

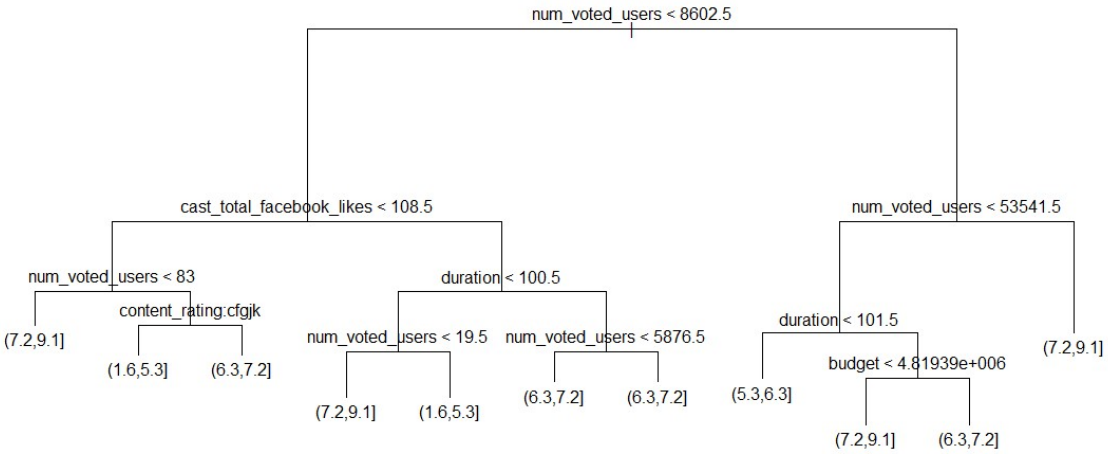
Appendix 1

Variable description:

Variable Name	Description
movie_title	Title of the Movie
duration	Duration in minutes
director_name	Name of the Director of the Movie
director_facebook_likes	Number of likes of the Director on his Facebook Page
actor_1_name	Primary actor starring in the movie
actor_1_facebook_likes	Number of likes of the Actor_1 on his/her Facebook Page
actor_2_name	Another actor starring in the movie
actor_2_facebook_likes	Number of likes of the Actor_2 on his/her Facebook Page
actor_3_name	Another actor starring in the movie
actor_3_facebook_likes	Number of likes of the Actor_3 on his/her Facebook Page
num_user_for_reviews	Number of users who gave a review
num_critic_for_reviews	Number of critical reviews on imdb
num_voted_users	Number of people who voted for the movie
cast_total_facebook_likes	Total number of facebook likes of the entire cast of the movie
movie_facebook_likes	Number of Facebook likes in the movie page
plot_keywords	Keywords describing the movie plot
facenumber_in_poster	Number of the actor who featured in the movie poster
color	Film colorization. 'Black and White' or 'Color'
genres	Film categorization like 'Animation', 'Comedy', 'Romance', 'Horror', 'Sci-Fi', 'Action', 'Family'
title_year	The year in which the movie is released (1916:2016)
language	English, Arabic, Chinese, French, German, Danish, Italian, Japanese etc
country	Country where the movie is produced
content_rating	Content rating of the movie
aspect_ratio	Aspect ratio the movie was made in
movie_imdb_link	IMDB link of the movie
gross	Gross earnings of the movie in Dollars
budget	Budget of the movie in Dollars
imdb_score	IMDB Score of the movie on IMDB

Graph 1. Variable Description (Taken from Kaggle.com)

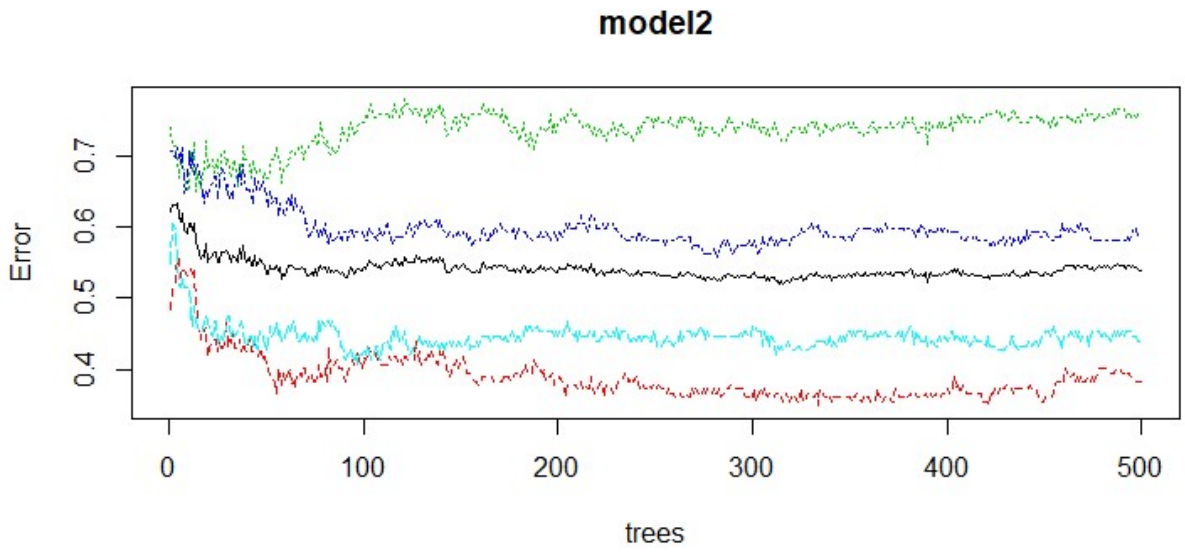
Appendix 2



Graph 2. Classification Tree for eliminated entries with missing "gross" values

	Variables	Importance	Rank
1	color	8.40	#14
2	num_critic_for_reviews	36.83	#5
3	duration	45.93	#3
4	director_facebook_likes	32.07	#10
5	actor_3_facebook_likes	31.60	#11
6	actor_1_facebook_likes	34.83	#7
7	num_voted_users	59.02	#1
8	cast_total_facebook_likes	35.53	#6
9	facenumber_in_poster	16.77	#13
10	num_user_for_reviews	46.64	#2
11	content_rating	25.08	#12
12	budget	37.53	#4
13	actor_2_facebook_likes	32.63	#8
14	movie_facebook_likes	32.33	#9

Graph 3. Random Forest Feature Importance Rank for eliminated entries with missing "gross" values



Graph 4. Random Forest clustering accuracy for eliminated entries with missing "gross" values

Special thanks to Professor Alex Belloni on his tremendous support, commitment and guidance on this project and this course throughout the semester.