National University of Singapore
Department of Statistics and Applied Probability
DSA4211 High-dimensional Statistical Analysis
Group Assignment

Group number:	4

	Matriculation of group members:
1.	A0157667M
2.	A0166737R
3.	A0157760Y
4.	A0154817Y

	Grade:
Introduction/Goals	
Model fitting	
Results	
Conclusion	
Penalize (Spelling, grammar, labels in plots, figures, tables, exceed page limit, format, late submission, plagiarism)	
Final grade	

Motivation

What makes a good movie? This question is quite subjective and I am sure the responses will vary among individuals as people have different taste and preferences. However, if we are able to collate and analyse all the different answers, we may be able to find a "model" answer to the question.

Goal

1. Find an interpretable model to predict public responses of movies accurately.
2. Find out the key attributes the majority of population thinks a good movie should have.

Description of Data

In order to collate and analyse all the different views people have on movies, we sourced a collection of movies' data[1] from the IMDB website. This data consists of 28 variables for 5043 movies from year 1905 to 2016. These 28 variables can be divided into 2 types. One that describes the different attributes of the movie and another that mainly quantifies the popularity of movie casts and directors. The descriptions of the 2 types of variables are shown in *Table 1* and *Table 2* below.

	Variables	**Description of Variable**		**Variables**	**Description of Variable**
1.	aspect_ratio	The aspect ratio the movie was made in	11.	movie_imdb_link	The movie's corresponding IMDB website link
2.	budget	The budget of the movie in dollars	12.	movie_facebook_likes	The number of facebook likes on the movie's page
3.	color	Coloured or black-and-white movie	13.	num_critic_for_reviews	The number of critic reviews for the movie
4.	country	The country where the movie is produced in	14.	num_user_for_reviews	The number of user reviews for the movie
5.	duration	The duration of the movie in minutes	15.	num_voted_users	The number of people who voted for the movie
6.	facenumber_in_poster	The number of actors featured in the movie poster	16.	plot_keywords	Keywords describing the plot of the movie
7.	genres	The genres the movie falls under	17.	title_year	The year in which the movie was released
8.	gross	Gross earnings of the movie in dollars	18.	language	Language of the movie
9.	imdb_score	IMDB Score of the movie	19.	content_rating	Content rating of the movie
10.	movie_title	Movie title			

Table 1: Description of variables that describes the different attribute of the movie

1 https://github.com/sundeepblue/movie_rating_prediction/blob/master/movie_metadata.csv

	Variables	Description of Variables
1	actor_1_name	The actor with the most facebook likes in the movie
2	actor_1_facebook_likes	The corresponding number of facebook likes of actor 1
3	actor_2_name	The actor with the 2nd highest facebook likes in the movie
4	actor_2_facebook_likes	The corresponding number of facebook likes of actor 2
5	actor_3_name	The actor with the 3rd highest facebook likes in the movie
6	actor_3_facebook_likes	The corresponding number of facebook likes of actor 3
7	cast_total_facebook_likes	The total number of facebook likes of the entire cast of the movie
8	director_name	The name of the director of the movie
9	director_facebook_likes	The number of likes on the director's facebook page

Table 2: Description of variables that quantify the popularity of casts and directors

Based on the data available, we feel that "imdb_score" will be a good indication of the public's perception on the quality of a movie. Hence, we will using "imdb_score" as the response variable and remaining variables as the predictors.

Data Pre-processing

Before we begin our analysis , it is important to perform data pre-processing in order to avoid getting misleading results and get a more accurate interpretation of our data.

Removal of trivial variables

We first remove trivial variables like the names of the directors and casts as we are more interested in how their popularity affects the "imdb_score" of a movie rather than just their names. The trivial variables removed are highlighted in *Table 1* and *Table 2*.

Removal of missing observations

We then removed observations with missing data in any of the predictors in order to have an unbiased analysis. This resulted in the removal of 1287 observations, however we still have 3756 observations which is still relatively large.

Creation of new variables

There are a few qualitative variables with a large number of different values. Using these variables directly will not be useful since it will result in a large number of subgroups with only a few observations in each of them. Therefore, we mitigate this problem by re-categorising them into smaller subgroups or converting them into quantitative variables. The description of the new variables is shown in *Table 3*.

Variable	**Description of variable**
ngenres	total number of genres the movie fall under
nkeywords	total number of plot keywords describing a movie
ccountry	the country the movie was produced in (“USA” , “UK” and “Others”
movie_age	Age of movie in year 2016 based on “title_year”

Table 3: Description of new variables

We converted “genres” into a quantitative variable as it allows us to investigate the ideal number of genres a movie should have. Similar to “genres”, we converted “plot_keywords” into a quantitative variable as it can tell us the complexity of a movie plot. Finding the ideal number of plot keywords allows us to get a sense of the ideal complexity a good movie plot should have. “ccountry” is a new variable which categorises the country the movie was produced in, into 3 subgroups, “USA”, “UK”, and “Others” . We converted “title_year” into “movie_age” as “title_year” merely represents the year the movie was released in. We are more interested to see if the public perceives newer movies to have better quality since they have better graphics as a result of the advancement in technology. After data pre-processing, we have a total of 21 predictors with 3756 observations.

Data exploration

There are various approaches that we can use to fit the data and their effectiveness largely depends on the characteristics of our data. As such, we plotted our response variable, “imdb_score” against of the predictors to have some sense of their relationship with the response variable before fitting appropriate models.

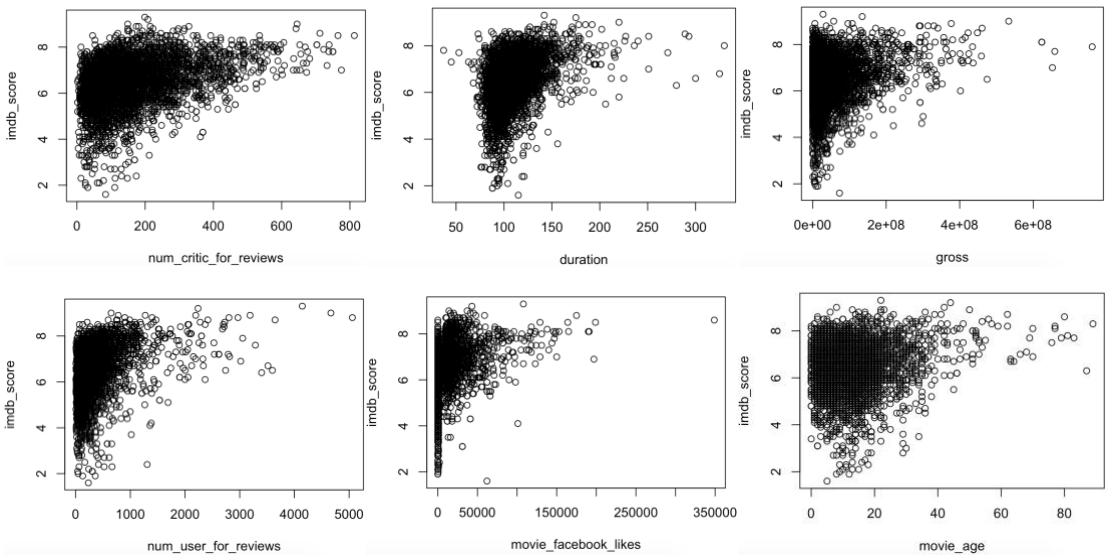

Figure 1: Scatter plots of some quantitative predictors with response variable

From the scatter plots in *Figure 1*, we can see that the predictors have a non-linear relationship with the response variable. This may give us a hint that we should be focusing on non-linear approaches rather than linear approaches to fit our data.

Model construction

Generally, our end goal is to get a model that is not only accurate in predicting the public's responses for movies , but also one that is interpretable. One very simple and interpretable model is multiple linear regression. We will be able to know the exact value of how each predictor influences the response variable through the values of the beta coefficients. However from our preliminary analysis, we found that a lot of the variables seem to have a non-linear relationship with the response variable. Therefore, a non-linear model may be less interpretable, however it may yield a much higher prediction accuracy.

We will first perform multiple linear regression to get a simple and interpretable model. Then we will perform boosting to try and get a non-linear model with much a higher prediction accuracy. We adopted a 80 : 20 training-test split for each method and evaluate each method using mean squared error (MSE) on the test set.

Multiple linear regression

The multiple linear regression model is a parametric model and assumes a linear relationship between the predictors and the response variable, "imdb_score". The best model fit is determined through the least squares approach - minimizing the error sum of squares, where each error term is the vertical deviations obtained from differencing each data point and the fitted line.

Before regressing the variables onto the response, we performed Forward Stepwise Selection and tune the variables to be selected in the model using the selection criteria - C_p, BIC and Adjusted R^2, and the obtained Cross Validation Errors. This enables us to obtain a parsimonious model, where the results obtained will be of better interpretability due to the presence of fewer variables in the model and prevents the presence of multicollinearity between the variables.

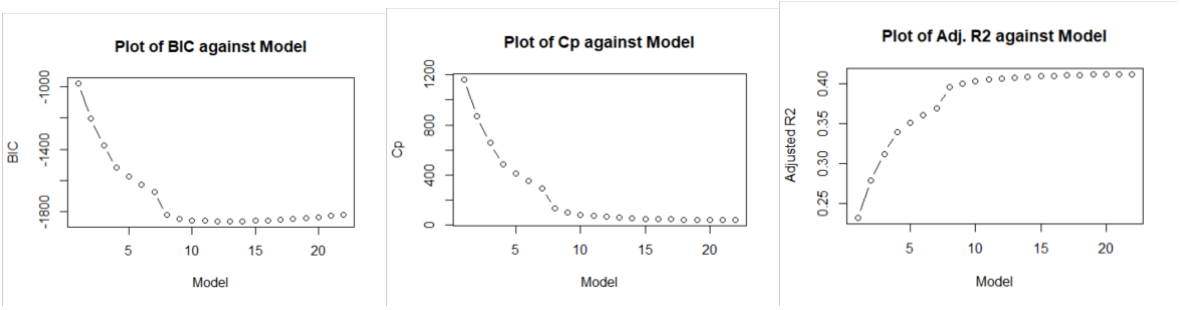

Figure 2: Plots of BIC, C_p, Adj. R^2 against Model

Figure 2 shows the plots of the criteria against the model. Based on the plots, we can see that the best parsimonious model we can choose via forward stepwise selection is the model with 9 variables (Model 9) as it is the simplest model with the lowest BIC, C_p and also the highest Adjusted R^2 value. The corresponding predictors in Model 9 is shown in Figure 3.

```
[1] "(Intercept)"          "num_critic_for_reviews" "duration"              "gross"
[5] "num_voted_users"      "num_user_for_reviews"   "content_ratingPG-13"   "movie_age"
[9] "ccountryUK"           "clanguageOthers"
```

Figure 3: Predictors in Model 9

From Figure 3, 'ccountryUK', 'clanguageOthers' and 'content_ratingPG-13' are factors of the different qualitative variables with multiple factor levels. Thus, it is equivalent to regressing on 'ccountry', 'clanguage' and 'content_rating', together with the other continuous variables.

We further performed feature selection based on Cross Validation Error, and check if the variables selected from Cross Validation are similar enough to the model selected earlier based on the 3 criteria, BIC, C_p and Adjusted R^2.

A 10-fold cross validation using the forward stepwise selection is performed to determine the number of variables to be selected for the regression model and the obtained cross validation plot is as shown in *Figure 4*.

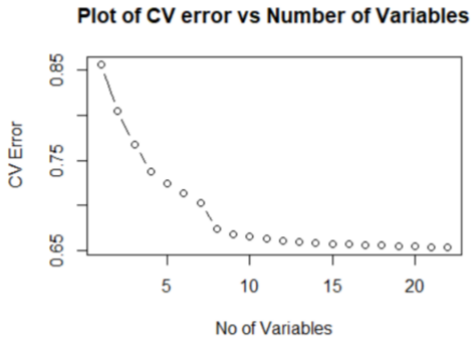

Figure 4: Plot of CV error against Number of Variables in the Model

Based on *Figure 4*, we observed that there is no significant decrease in the cross-validation errors when more than 9 predictors are selected in the model. This is in line with the conclusion obtained using the selection criteria - selecting Model 9. Therefore, we will use the 9 predictors determined by the forward stepwise selection for our regression model. Additionally, by performing forward stepwise selection, these predictors identified are considered to be the most important predictors in determining the IMDB scores.

The refitted model based on cross validation with forward stepwise selection is shown below:

$$\widehat{imdb\_score} = \hat{\beta}_0 + \hat{\beta}_1(num\_critic\_for\_reviews) + \hat{\beta}_2(duration) + \hat{\beta}_3(gross) + \hat{\beta}_4(num\_voted\_users) + \hat{\beta}_5(num\_user\_for\_reviews) + \hat{\beta}_6(content\_rating) + \hat{\beta}_7(movie\_age) + \hat{\beta}_8(country) + \hat{\beta}_9(clanguage)$$

The corresponding coefficients for the 9 predictors selected are given in *Figure 5 below*.

(Intercept)	num_critic_for_reviews	duration	gross
4.686647e+00	2.238748e-03	9.474131e-03	-1.463883e-09
num_voted_users	num_user_for_reviews	content_ratingPG-13	movie_age
3.510008e-06	-5.401066e-04	-2.662396e-01	2.108931e-02
ccountryUK	clanguageOthers		
3.237690e-01	8.906347e-01		

Figure 5: Estimates of Coefficients for the selected regression model

From *Figure 5*, the estimates of coefficients obtained are equal to the change in IMDB movie scores associated with the corresponding predictor, holding the other predictors fixed. They give us a sense of the importance of each predictor and how each predictor influences IMDB movie scores individually. Predictors with positive coefficients will have a positive linear relationship with IMDB movie scores and vice versa.

For example, increasing the duration of the movie by 1 minute while keeping the other predictors constant, will lead to an approximate 0.00947 increase in its IMDB score. On the other hand, if the number of user reviewing the movie increases by 1 while the other predictors are kept constant, it will lead to a 0.000540 decrease in IMDB score.

We then fit the test data into this model and obtained a test MSE of 0.717, which will be used for the comparison between methods later in the report.

Boosting

The other method that we explored is boosting. Boosting is a form of tree-based method. Tree-based methods segment the predictors space into simple regions and the prediction for an observation is taken to be the mean of training observations in the region it belongs. Boosting aims to improve the prediction accuracy by using and combining information from previously grown trees and building the new trees sequentially. Although boosting generally produce a less interpretable model as compared to multiple linear regression, boosting may have a much higher level of prediction accuracy.

The predictors used in multiple regression are selected using parametric methods, however, tree-based methods are non-parametric, therefore they may not be suitable. Thus, we fit the model using all the predictors in boosting. To fit the model into the data, we performed 10-fold cross-validation to select the number of decision trees *B,* to avoid overfitting and their interaction depth *d,* to control the complexity of the boosted ensemble while fixing shrinkage parameter to be 0.1. Small values of shrinkage parameter can require very large *B* to achieve good performance. Hence, we decided to fix shrinkage to be 0.1 instead of tuning it in order to cut down on computational cost.

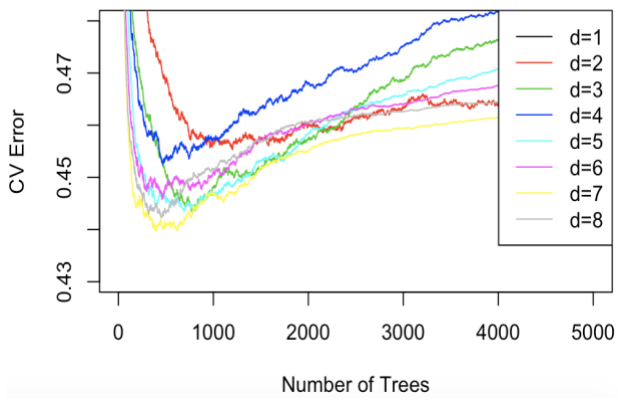

Figure 6: Cross-validation plot for tuning B and d

Figure 6 shows the cross-validation plot to choose the tuning parameters B and d. The different coloured lines represent the different interaction depths d used. We can see that the yellow line which corresponds to d = 7 interaction depth gives the lowest CV error of 0.440 at B = 400 number of trees. Therefore, 400 trees of 7 interaction depth are used for our boosting model and yields a test MSE of 0.432.

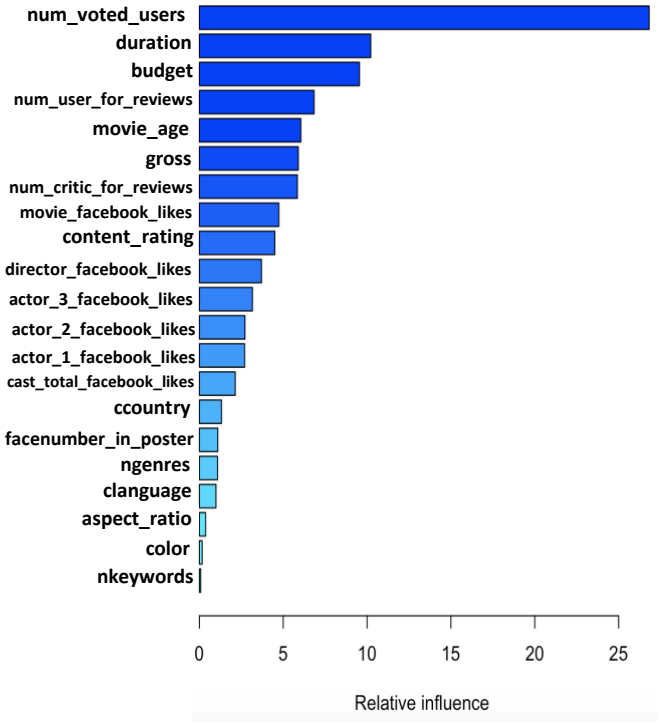

Figure 7: Variable Importance plot

Boosting aims to improve prediction accuracy by combining decision trees, this will lead to some loss of interpretability in the model. However, we can still use the total Residual Sum of Squares decreased due to splits over a predictor to determine the relative importance of the predictors. Dependence plots can be used afterwards to get a sense of how the important predictors influence the response variable.

Figure 7 shows the variable importance plot for our boosting model and we can see that "num_voted_users", "duration" and "budget" is the top three most important predictors.

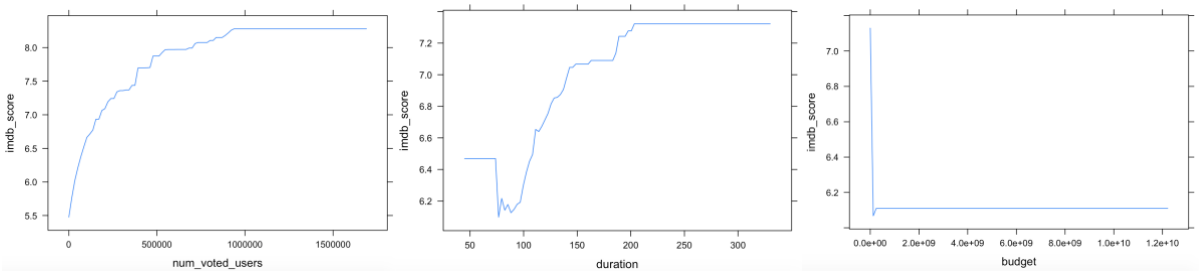

Figure 8: Dependence plots of the top 3 most important variables

From *Figure 8*, we can see that both "num_voted_users" and "duration" generally have a positive relationship with "imdb_score". However, both plots reached a stagnant after achieving certain values. Therefore, movies should strive to reach these values for "num_voted_users" and "duration" first, before working on other aspects of the movie. Getting the optimal values for these two predictors will be most efficient way in increasing the "imdb_score" of a movie.

On the other hand, "budget" has a pretty weird relationship with "imdb_score" since it starts with a negative relationship. It seems that people are generally more impressed if movie producers are able to work with a very low budget compared to an average or high budget.

Comparison of Results

We will be evaluating the performances of the two models based on the their respective test MSE values.

Method	Multiple Linear Regression	Boosting
Test MSE	0.717	0.432

Table 4: Performances of the two models used

From *Table 4*, we can see that our boosting model yields a much lower test MSE and hence, it has a much better performance compared to the multiple linear regression model.

This disparity in performances may be due to the evidences of non-linear relationship that the quantitative predictors have with "imdb_score", which was found earlier in the preliminary analysis. In order to investigate further, we plotted a residual plot for the multiple linear regression model.

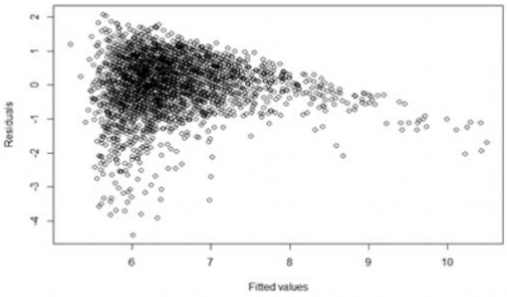

Figure 9: Plot of Residuals against Fitted Values

A residual plot is suitable in helping to identify nonlinearity in the dataset. If a linear trend can be seen in the plot, it suggests that the linear regression model is a good fit for the dataset. However, based on *Figure 9*, we can see that there is an obvious quadratic trend, and this further supports the evidences of a non-linear association that the predictors have with "imdb_score". Therefore, non-linear models, such as boosting, may be more appropriate and thus yielded a better performance compared to multiple linear regression model.

Besides the test MSE, the models seem to have a severe contradiction on the importance of the variables. The importance of each predictors in the multiple linear regression model is based on the absolute coefficient values.

	Multiple Linear Regression	**Boosting**
1.	clanguage	num_voted_users
2.	ccountry	duration
3.	content_rating	budget

Table 5: Top 3 important variables for the two models

From *Table 5*, we can see that the top 3 most important variables for the 2 methods greatly differ. Furthermore from *Figure 7*, "clanguage" is one of the least important variable in boosting model. This may be due to the strength of the non-linear relationship the predictors have with "imdb_score". The predictors may have a weak linear relationship but very strong non-linear relationship with "imdb_score". Therefore, such predictors are not chosen in the forward stepwise selection.

Conclusion and Future Improvements

In conclusion, we feel that boosting is the better model to fulfill our goals. Even though the model is less interpretable as compared to multiple linear regression model, we are still able to identify the key attributes a good movie should have through the variable importance and dependence plot. At the expense of some interpretability, boosting have a much higher accuracy compared to multiple linear regression model.

However, there is a huge limitation to our current model. We will not be able to predict the "imdb_score" for upcoming movies due to the missing values for important predictors such as "num_voted_users", since they will only be available in the future when the movie is released. Currently, our only solution is to estimate these missing values using the mean or median. These are poor estimates and they will thus limit the accuracy of predictions.

Therefore, for future improvements, we can work on finding models to predict these missing values, so we can plug them into our boosting model to predict the "imdb_score" of upcoming movies more accurately.