Movie Success Prediction

PROJECT REPORT

Rakesh Parappa | U01382090 | CS660

Abstract

The report entails analyzing different variables like movie budget, actor's Facebook likes, director's Facebook likes and others to determine the relationship in historical data. Below are the initial questions answered in the report:

- Whether the IMDB rating has the positive role in movie's success.
- Whether the popularity of the actor as any impact on movie's success.
- Whether the critic reviews have any adverse effect on the movie's success.

Finally, the report has models built to determine the success of the movie. Models used are linear regression, logistic regression, decision trees and random forest.

Introduction

Every year there are hundreds and thousands of movies released in a different genre. Some have big Hollywood stars and some have new. Sometimes movies with big budget perform badly at box-office and movies with low budget performs exceptionally well.

There are many rating websites like IMDB which rates the movies on various factors and are widely considered by users before watching a movie. These ratings sometimes determine the success of the movie.

The dataset required for the project is obtained from kaggle's website. The dataset had 5000 movie records and 28 variables or attributes. Among the 28 variables, 16 key variables are considered:

```
 [1] "color"                 "num_critic_for_reviews"  "duration"
 [4] "facenumber_in_poster"  "director_facebook_likes" "actor_3_facebook_likes"
 [7] "actor_2_facebook_likes" "actor_1_facebook_likes"  "gross"
[10] "num_voted_users"       "genres"                  "country"
[13] "budget"                "title_year"              "imdb_score"
[16] "result"
```

Data Preparation

The variables "color", "country" and "title_year" are converted to factor. All the movies with color as "na" were changed to either "Color" or "Black and White" based on the below rules:

```
movies.req$color[is.na(movies.req$color) & movies.req$title_year>2000] <-
"Color"
```

```
movies.req$color[is.na(movies.req$color)] <- "Black and White"
```

The “result” is a binary variable manually created with values “yes” and “No” to signify movie success. Below are the rules used to create the variable “result”:

```
movies.req$result[movies.req$budget>movies.req$gross] <- 1
movies.req$result[movies.req$budget<=movies.req$gross] <- 2
movies.req$result<- factor(movies.req$result,levels = c(1,2),labels =
c("No","Yes"))
```

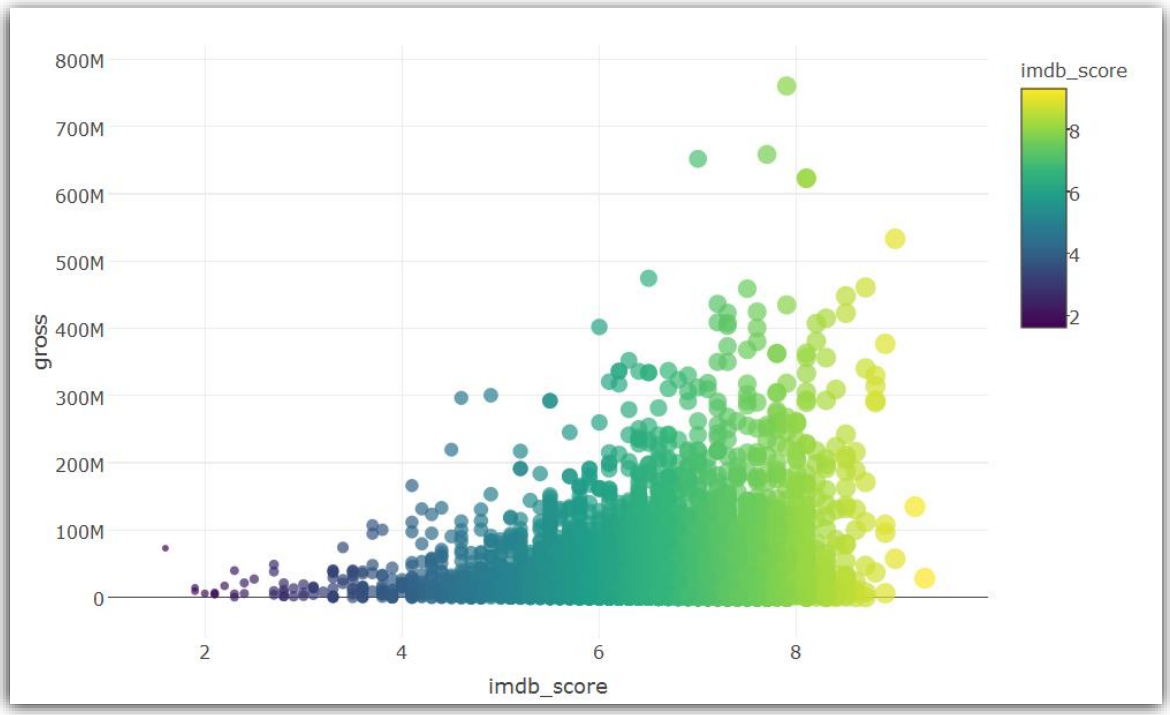
All other records which have no “budget” and “gross” are omitted since our conclusion of movies success is solely based on the above variables.

With all the omission, the final dimension of data considered is 3873 records and 16 variables.

Exploratory Data Analysis

Gross VS IMDB score

Many people watch movies based on the IMDB score. Below diagram shows the relationship between IMDB score and the gross of the movie.

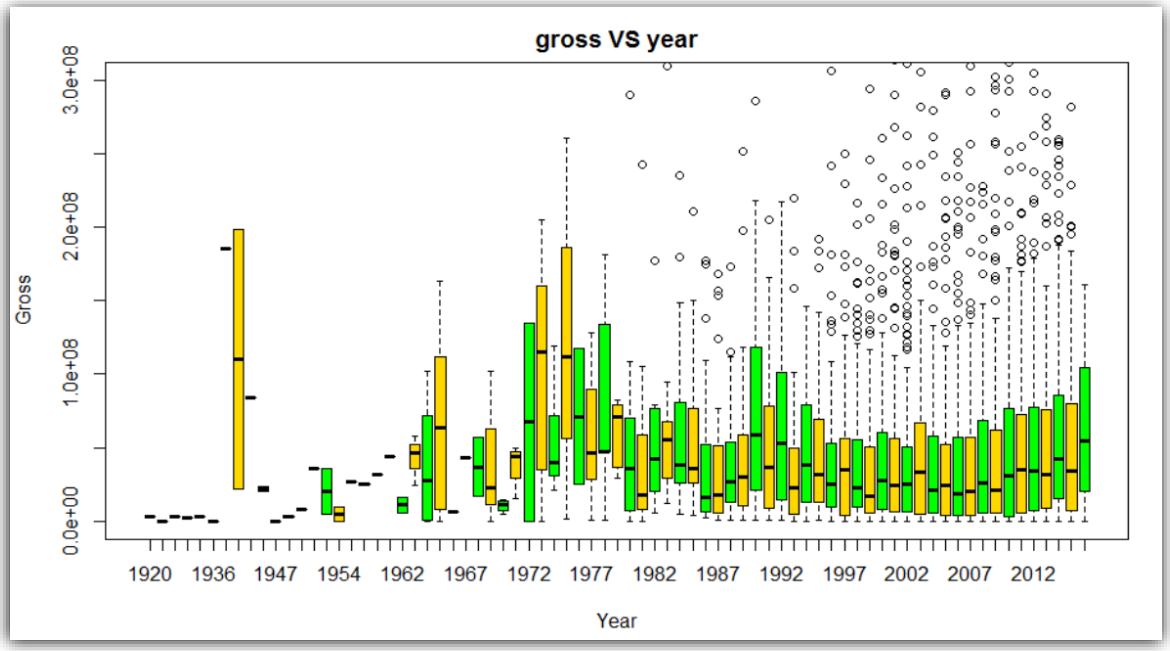


Movies with rating > 8 are considered to be very good movies. Movies with rating 7 to 8 are considered to be good movies and movies with rating 1 to 5 are the movies that one

should avoid watching. In the above plot, we can see that approximately till IMDB rating 7.5, an increase in the rating, increases gross. Post that rating tends to have a negative effect on the gross. Usually, IMDB score > 7.5 are critically acclaimed movies which might not perform well at box office.

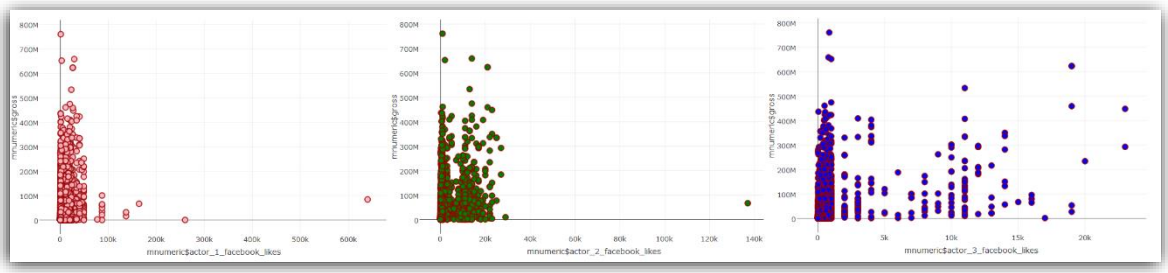
Gross VS Year

The gross of the movies has abnormal distribution until the year 1980 since there were less number of movies being made until then. Post that the gross per year seems to be almost same, having a mean of around 50M. There are many outliers as some movies performed exceptionally well. Below diagram shows the barplot of distribution of gross over the years. Gross is scaled to 300M to show the distribution and there are many movies having gross more than that.



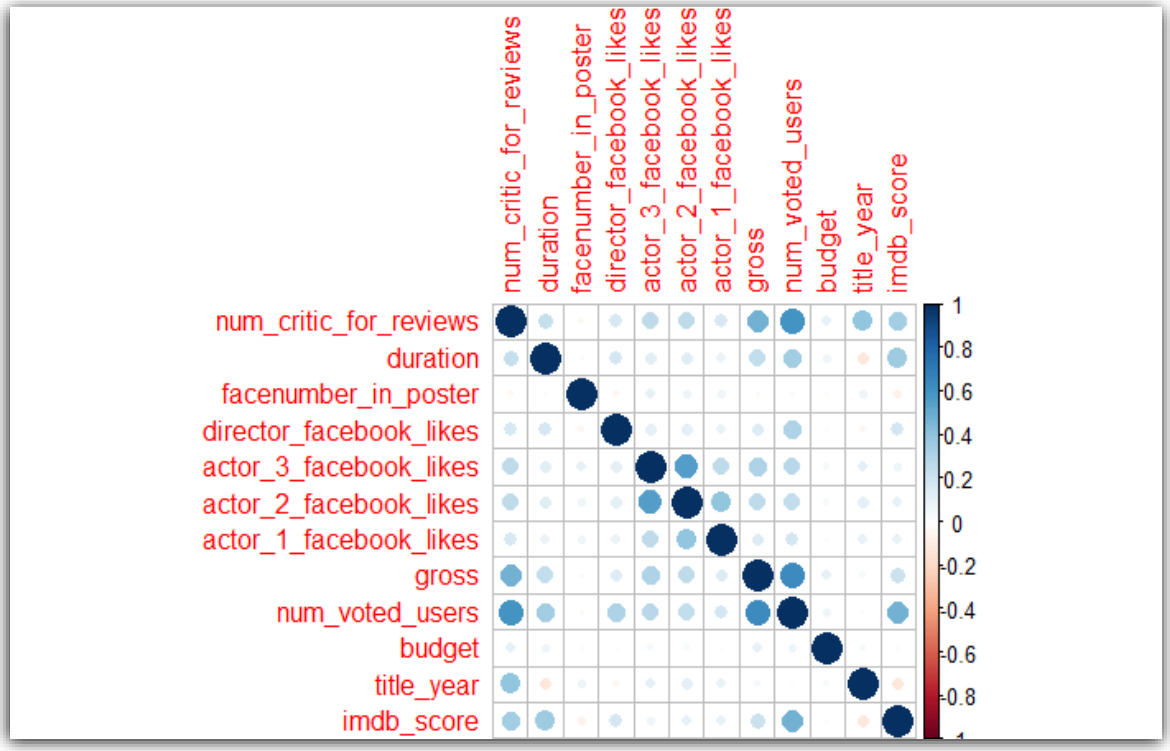
Gross VS Actors Facebook Likes

Facebook likes of actor 1 is greater than Facebook likes of actor 2 and actor 3 has lowest Facebook likes. But the amount of Facebook likes doesn't seem to have any different effect on the gross as all the three actor plots have movies with higher as well as lower gross. There seems to be no positive or negative relationship as well to show that increase in Facebook likes has any effect on gross.



Correlation

Below is the correlation matrix of 12 numeric variables.



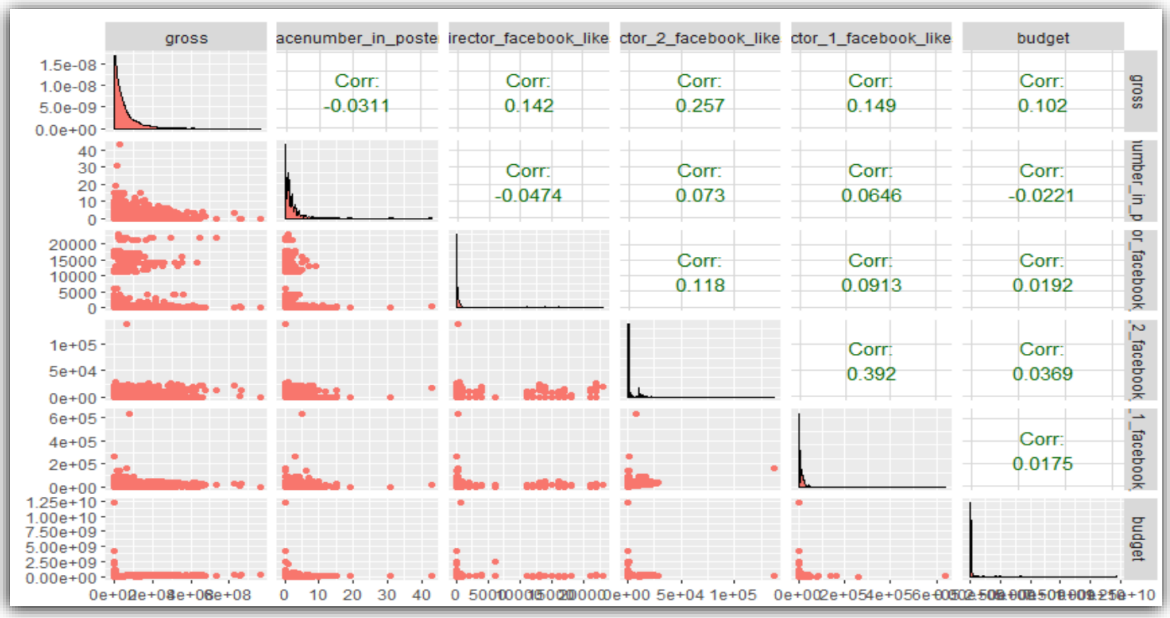
Below are some of the insights from the correlation:

- Gross has a higher correlation with the “num_voted_users” and “num_ critic_for_reviews”.
- Gross seems to have low correlation with “IMDB_score”. A higher score doesn’t mean that movie has done well.
- Among all the actors Facebook likes, “actor_3_facebook_likes” has a higher correlation with the gross.

- Gross seems to have very low correlation with the “budget”.

To be more clear below is the pairwise correlation matrix of some of the positively correlated variables.

We can verify that all the above insights made seems to be right. The above graph also shows the effect of one variable over the other. The highest correlation of 0.63 seems to be between “num_voted_users” and “IMDB_score” which probably means movies with higher voted users have a higher score or probably the popular movies have a higher score. Below is the relationship graph of other variables which has minimal impact on each other.



Build and Validate Model

After the initial analysis, various insights and associations between the variables were found. Since most of the variables are numeric, it is easy to build a linear regression model.

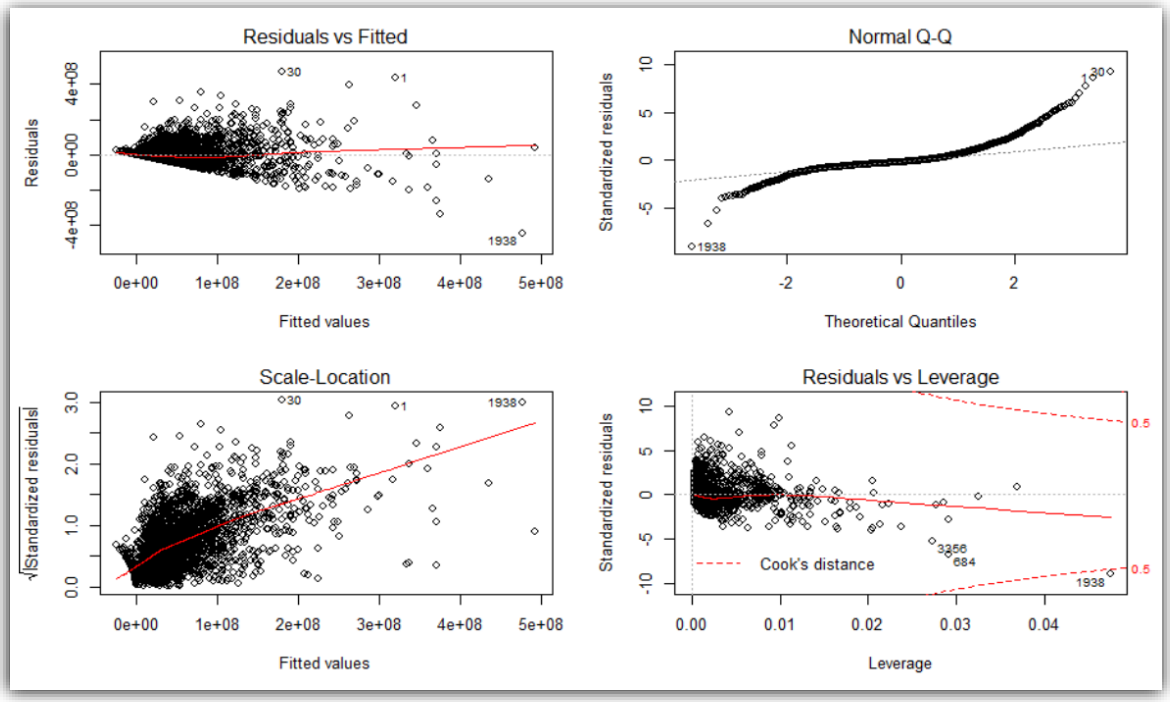
The goal is to fit a predictive model to an observed data set of dependent i.e. gross, and exploratory variables. After developing a model if we use exploratory variables on it, then it should predict the value of gross with minimum error.

Below is the summary of an initial linear regression model with only significant variables. All other exploratory variables whose p-value was greater than 0.05 were removed from the model.

```
Call:
lm(formula = gross ~ color + num_critic_for_reviews + duration +
    director_facebook_likes + num_voted_users + imdb_score, data = movies.req)

Residuals:
       Min          1Q      Median          3Q         Max
-447802494  -22677835   -9169966   13351974  472776057

Coefficients:
                          Estimate   Std. Error t value      Pr(>|t|)
(Intercept)           25064232.898  7786210.930   3.219        0.0013 **
colorColor            17613264.428  4457310.617   3.952 0.0000789411 ***
num_critic_for_reviews    88061.702     8174.142  10.773       < 2e-16 ***
duration                214481.461    38992.249   5.501 0.0000000401 ***
director_facebook_likes   -1122.908      284.524  -3.947 0.0000805732 ***
num_voted_users             276.840        7.464  37.088       < 2e-16 ***
imdb_score                -8850291.625   883407.036 -10.018       < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51240000 on 4144 degrees of freedom
  (892 observations deleted due to missingness)
Multiple R-squared:  0.4409,    Adjusted R-squared:  0.4401
F-statistic: 544.8 on 6 and 4144 DF,  p-value: < 2.2e-16
```

Let’s look at the residual plots to determine the quality of the fitted model.

1. *Residual vs Fitted Plot*: The residuals should be equally spread across the horizontal line or zero mean without a distinct pattern. The points seem to be spread across the horizontal line, however, there seems to be a pattern where the points tend to spread more with the increase in the fitted values.
2. *Q-Q Plot*: The plot shows whether the data residuals are normally distributed. The plot has some points having a deviation in the beginning and as well in the end. Hence the normality assumption fails.
3. *Scale-Location*: This plot checks for the assumption of homoscedasticity i.e. if the residuals have equal variance. The points have a pattern where the variance tends to increase with the increase in the fitted values. Hence the assumption fails.
4. *Residuals vs Leverage*: The plot identifies outliers with high leverage points with high influence. There are points with high leverage but one record 1938 is below cook’s distance and hence have high influence.



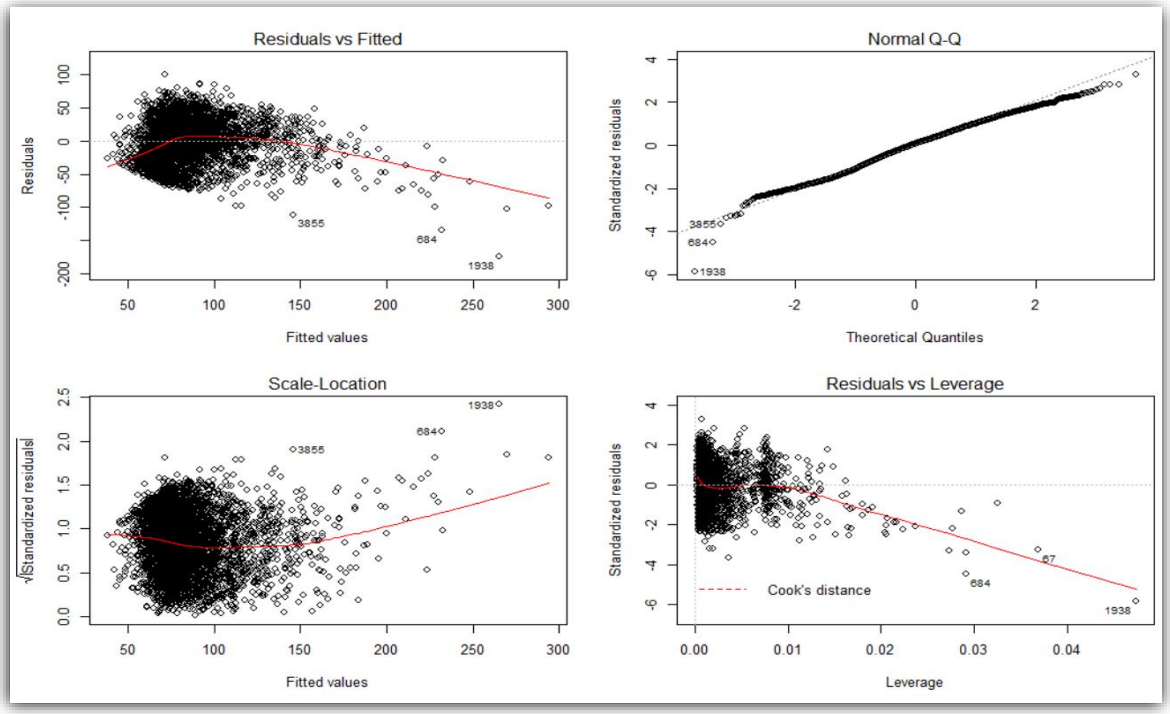
To fix the above model, Box-Cox transformation is used. The statisticians George Box and David Cox developed a procedure to identify an appropriate exponent (Lambda = l) to use to transform data into a "normal shape." The Lambda value indicates the power to which all data should be raised. In order to do this, the Box-Cox power transformation searches from Lambda = -5 to Lamba = +5 until the best value is found. Below is the transformation formula:

$$T(Y) = Y^{l}$$

After applying the Box-Cox transformation, below is the refitted model.

```
Call:
lm(formula = gross.boxc ~ color + num_critic_for_reviews + duration +
    num_voted_users + imdb_score, data = movies.req)

Residuals:
     Min       1Q   Median       3Q      Max
-172.459  -20.824    2.488   22.003  100.226

Coefficients:
                        Estimate  Std. Error t value Pr(>|t|)
(Intercept)           76.44588437  4.62919699  16.514  < 2e-16 ***
colorColor            13.13461839  2.65803764   4.941 8.06e-07 ***
num_critic_for_reviews  0.07272339  0.00487779  14.909  < 2e-16 ***
duration                0.16928658  0.02283375   7.414 1.48e-13 ***
num_voted_users         0.00011846  0.00000436  27.167  < 2e-16 ***
imdb_score             -7.00973051  0.52508516 -13.350  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.59 on 4148 degrees of freedom
  (889 observations deleted due to missingness)
Multiple R-squared:  0.3614,    Adjusted R-squared:  0.3606
F-statistic: 469.5 on 5 and 4148 DF,  p-value: < 2.2e-16
```

"director_facebook_likes" is removed from the above model as it was not significant in the refitted model. Below are the residual plots.



The transformation seems to have fixed the model and all the assumptions of linear regression are satisfied.

Now the model is validated with the test data, and below are the results.

duration	movie_title	imdb_score	budget	gross	predicted
178	Avatar	7.9	237000000	760505847	857883877
169	Pirates of the Caribbean: At World's End	7.1	300000000	309404152	175191673
148	Spectre	6.8	245000000	200074175	162855303
164	The Dark Knight Rises	8.5	250000000	448130642	1401908057
132	John Carter	6.6	263700000	73058679	94713634
156	Spider-Man 3	6.2	258000000	336530303	176147133
100	Tangled	7.8	260000000	200807262	59950572
141	Avengers: Age of Ultron	7.5	250000000	458991599	258093931
153	Harry Potter and the Half-Blood Prince	7.5	250000000	301956980	106532687
183	Batman v Superman: Dawn of Justice	6.9	250000000	330249062	278682314
169	Superman Returns	6.1	209000000	200069408	131041440
106	Quantum of Solace	6.7	200000000	168368427	108931284
151	Pirates of the Caribbean: Dead Man's Chest	7.3	225000000	423032628	186258895

Since the goal was to determine the success of the movie, the variable "result" is used as the class with binary values "yes" and "No". Different classification models were fitted and below are their results.

Data is divided into 80% training and 20% test data.

Logistic Regression

	Predicted	
Actual	No	Yes
No	273	104
Yes	116	282

Decision Tree

	Predicted	
Actual	No	Yes
No	245	132
Yes	99	299

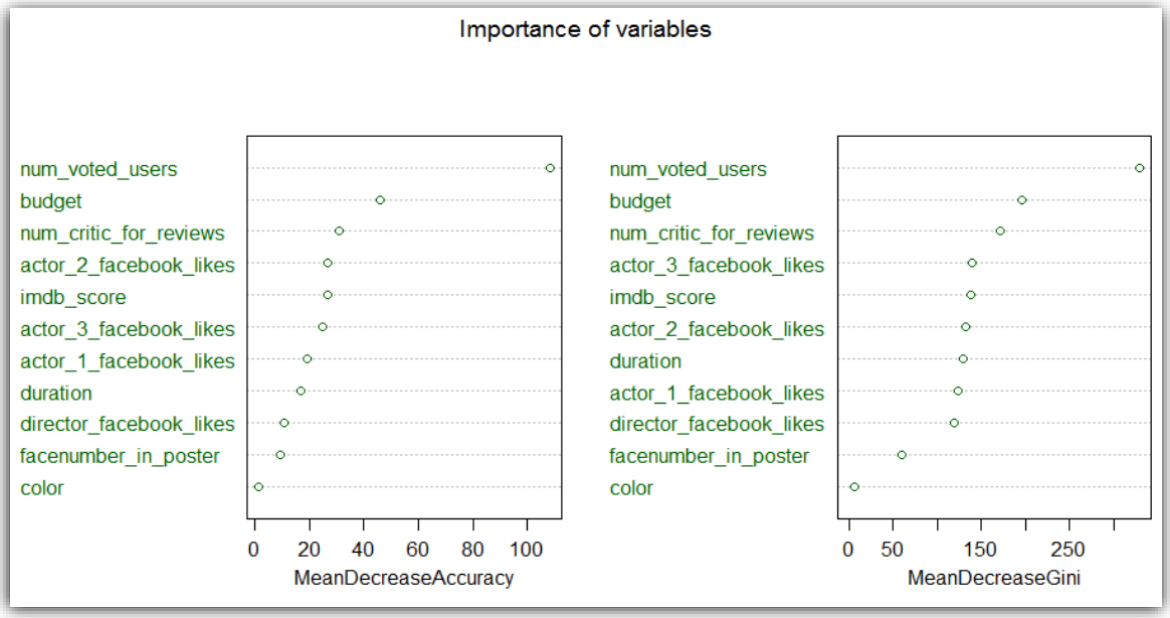
Random Forest

	Predicted	
Actual	No	Yes
No	262	115
Yes	79	319

Below is the performance metrics. Random Forest outperforms the other two classification models with the accuracy of 75% but the true negative rate is 69% which means out of all non-successful movies 69% records classified correctly as non-successful. Below are the importance plots of different variables according to the random forest model.

Performance Metrics			
	Logistic Regression	Decision Tree	Random Forest
Sensitivity	0.71	0.75	0.8
Specificity	0.72	0.65	0.69
Pos Predictive value	0.73	0.69	0.74
Neg Predictive value	0.7	0.71	0.77
Accuracy	0.72	0.7	0.75

Based on the random forest model, "num_voted_users" is very important variable to determine the success. "Budget" and "num_critic_for_reviews" are the next important variables. Surprisingly, all three actors Facebook likes are also important to determine the result of the movie. Of all the three "actor_2_facebook_likes" is the most important variable.



Conclusion

Based on the linear regression model used to predict the gross and random forest model used to classify the result of the movie, below is the conclusion:

- The most important variable to get the higher gross is duration. Which means, longer the duration, higher will be the gross of the movie. Probably that is the reason why movies like "Lord of the rings" worked so well in box office.
- Based on the random forest, a number of voted users has the highest importance, i.e. higher the number of voted users, higher the probability of success.
- Budget is deemed important variable in the random forest and based on the regression model, higher the budget, lesser the gross. So spend wisely.
- IMDB rating has an inverse impact on the gross. Higher the IMDB rating, lesser will be the gross. Probably, the higher rating movies (greater than 8) are critically acclaimed with low box-office collection.
- Number of critic reviews also important. Higher the number of reviews, higher the movie success.

Further improvement can be done on the models by considering movies based on country and genre.

References

https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

http://data.library.virginia.edu/diagnostic-plots/

https://www.isixsigma.com/tools-templates/normality/making-data-normal-using-box-cox-power-transformation/

http://stackoverflow.com/questions/33999512/how-to-use-the-box-cox-power-transformation-in-r

http://www.listendata.com/2014/11/random-forest-with-r.html