Predicting IMBD Scores

Shivani Dedhia, Akhila Pamukuntla, Nafis Chowdhury, Akshita Jain

STA 9750 Final Project

Introduction

Many factors make a movie successful and profitable. Some of the factors include budget of the movie, actor's and director's popularity, etc. We will find out which factor has the most impact on a movie's success and profitability.

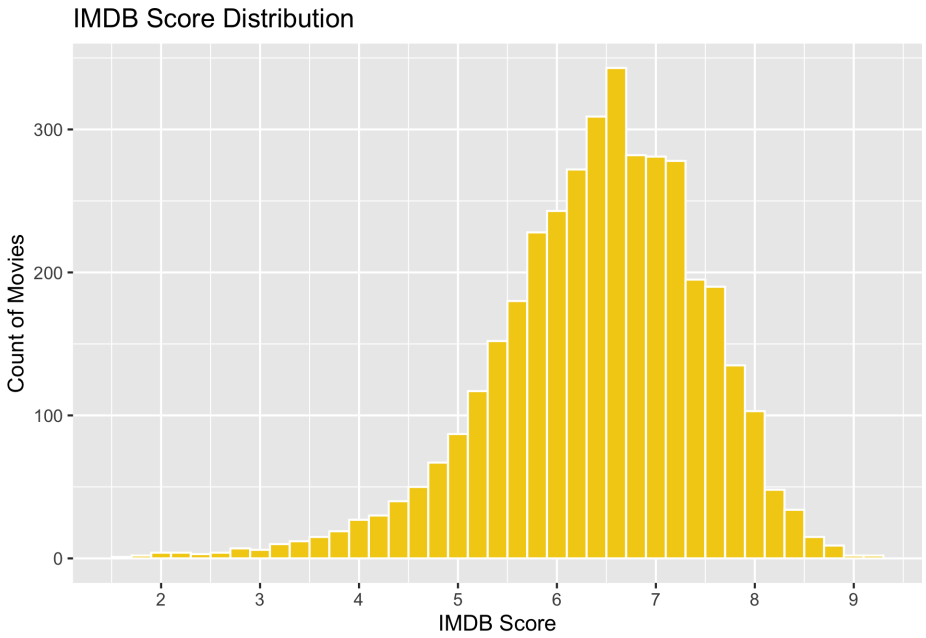
We have the IMDB 5000 movie dataset (https://www.kaggle.com/suchitgupta60/imdb-data), which has 28 variables, 5043 movies across 100 years in 66 countries. There are many variables in the dataset such as Director, Actors, Duration, Gross, Budget, Genres, Facebook Likes, etc.

We will be using some of the modeling techniques with associated visualizations to identify the most important variables that impact the success and ratings of the movie along with the profitability.

Data Exploration

After exploring the data variables, languages, aspect ratio, imdb movie link and color, we removed the 4 variables to reduce bias in the data. These variables had data heavily skewed on one side. Movies rated above 7.5 are considered to be highly recommended. Majority of the movies are rated 7.6 with only a handful rated above 9. The highest rating received by one movie is 9.3. This shows there is not enough data to understand the variables that have the strongest impact on highly rated movies.

IMDB offers a scoring scale that allows users to rate films on a scale of one to ten. It indicates that submitted scores are filtered and weighted in various ways in order to produce a weighted mean that is displayed for each film, series, and so on. So, this means according to this, the higher the score is, the better the movie will be. So, here most of the movies are between the range of 6.5 to 7.7. We can consider this as good/very good movies. It indicates, the acceptance of these movies are well enough among the movie audiences. However, there are obviously some exceptionally phenomenal movies which are rated beyond 8 but we can see that numbers are may be couple of hundreds only. On the other side, the bar chart is very skewed which indicates there are also a good number of movies which were upto the expectation of the viewers. However, in general, this is still difficult to predict the important factors behind these scores.




```
## # A tibble: 17 x 2
## # Groups:   imdb_score [17]
##    imdb_score     n
##         <dbl> <int>
##  1        7.6   100
##  2        7.7    90
##  3        7.8    83
##  4        8      55
##  5        7.9    52
##  6        8.1    48
##  7        8.2    24
##  8        8.3    24
##  9        8.5    19
## 10        8.4    15
## 11        8.6     8
## 12        8.7     7
## 13        8.8     5
## 14        8.9     4
## 15        9       2
## 16        9.2     1
## 17        9.3     1
```

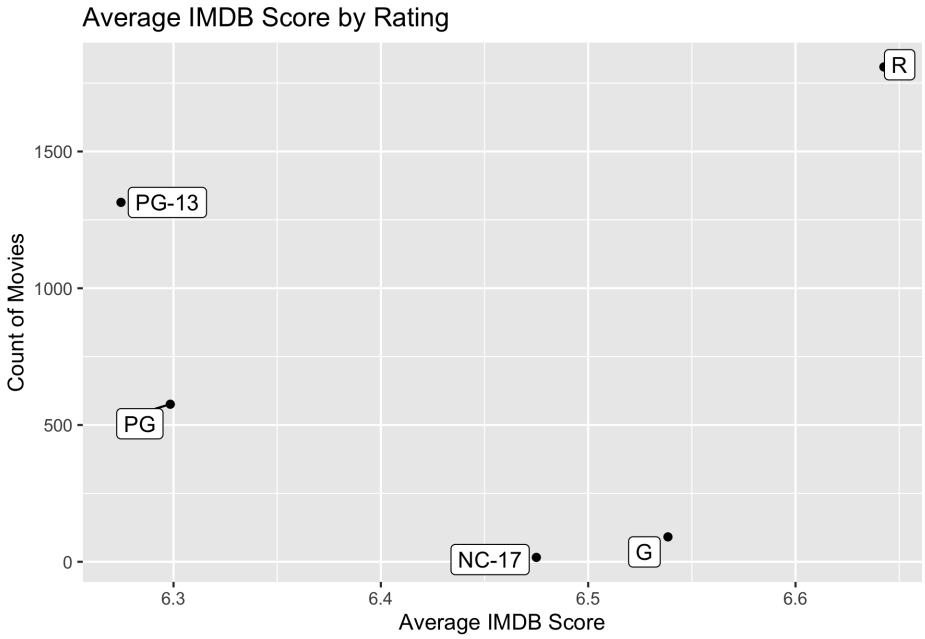
Impact of content rating on imdb score

Average score for movies by content rating is below 6.6 which is considered a poor imdb score. As per this distribution, content rating does not have a strong impact on the imdb score.

Content rating basically tells us about the age group of people who should watch these movies. Content rating does not have strong influence on the imdb score. However, we can see that most of the content ratings are in the range of 6.5 to 7.5. But, there are some outlier categories which are beyond this range. So, it means there is a possibility that these ratings are coming from different age groups but we cannot confirm this certainly with this data. R and PG-13 have the most number of movies which indicates most of

the movies are not for under age people. What makes it more interesting is the imdb score of these movies are also very poor and below average. To summarize, a big chunk of the movies are the restricted ones with poor imdb score.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

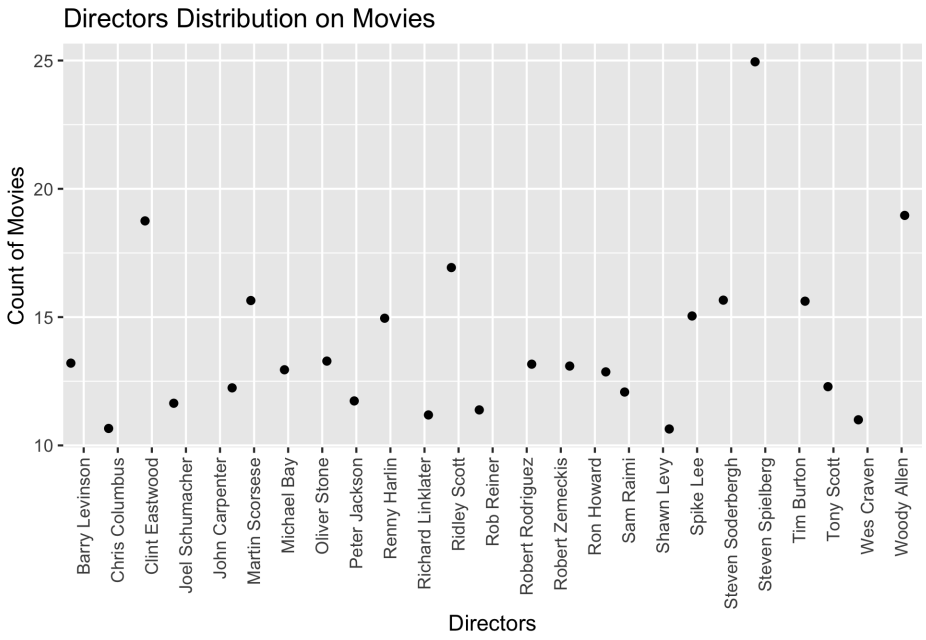


Understanding the distribution of directors and their effect on IMDB score

Directors have been grouped by the number of movies directed. The data has been filtered to only show movies directed above 10 and below 50 to remove any anomalies in the data. Directors with more movies could have a higher fan following, credibility and success rate possibly leading to a higher imdb score.

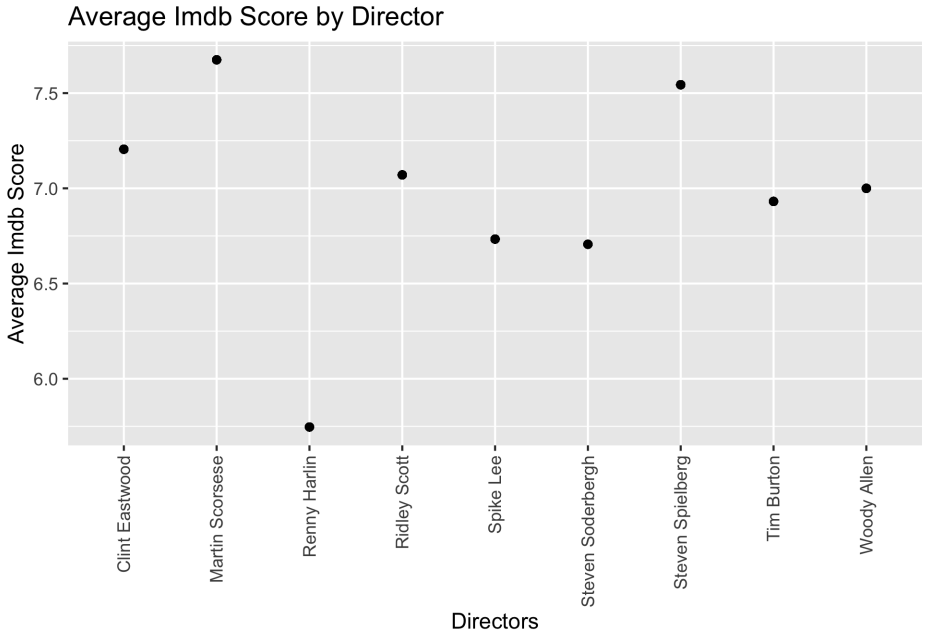
According to the below distribution, even after filtering, the number of movies for most of the directors are between 10 to 15, few are in the range of 15 to 20 and rest two are absolutely outliers. This indicates that in this time frame, the most naturalistic production of movies by the directors are between 10 to 15 range. The rational can be budget, resources or time constraints etc.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



Only a handful of directors have over 15 movies directed in this data set. Steven Spielberg is the only director to have ~ 24 movies directed. The chart below shows the average imdb score for directors with 15 or more directed movies. The imdb score is above 5.5 for directors with more than 15 movies. Most directors have received a higher imdb score that shows the number of movies directed has a slight impact on the imdb score.

In the below chart, we are observing the average imdb score for the directors who has movies between 15 to 49. We see most of the averages lies between 6.75 to 7.75. This tells us the average of the score has similar trend for this group of directors. However, there is an outlier as well.

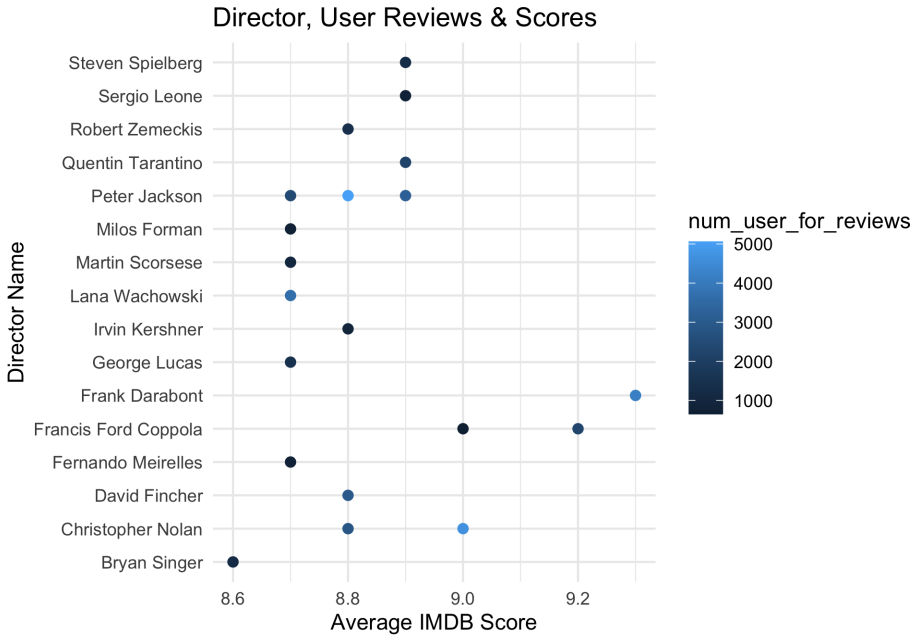


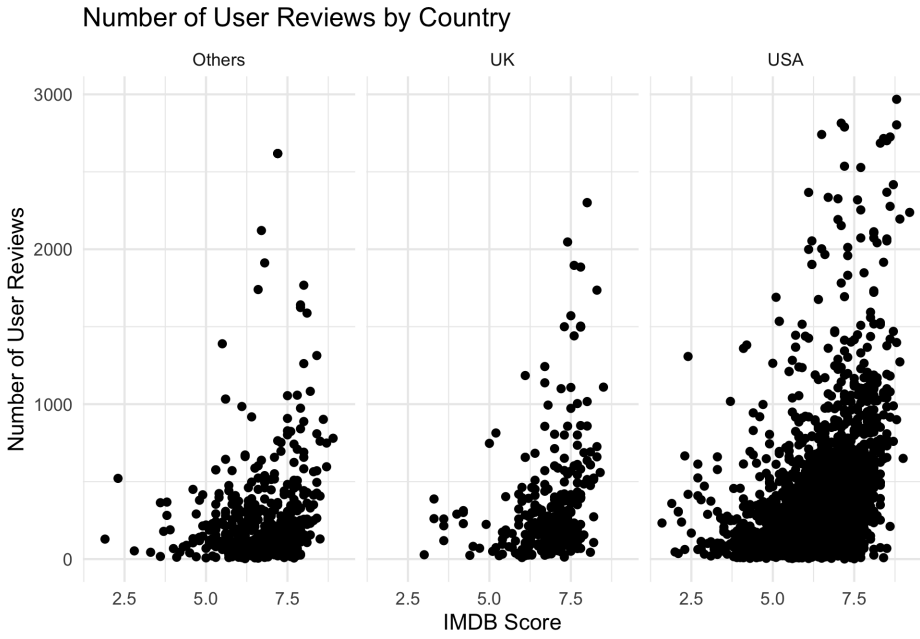
Impact of movie duration for a higher imdb score

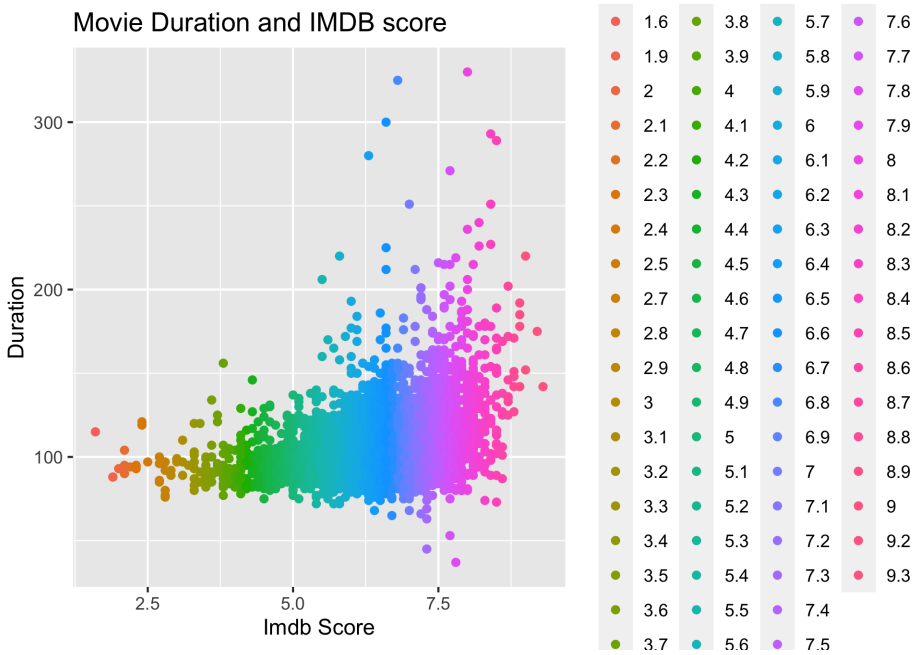
The average imdb score is shown compared to the duration of the movie. The duration of the movie increases as the imdb score increases. Duration of the movie seems like it is one of the impactful factors that affect the imdb score. The number of movies rated above 7 increases as the movie duration increases.

Top 20 movies grouped by genres with the highest imdb score.

```
## `summarise()` regrouping output by 'director_name' (override with `.groups` argument)
```





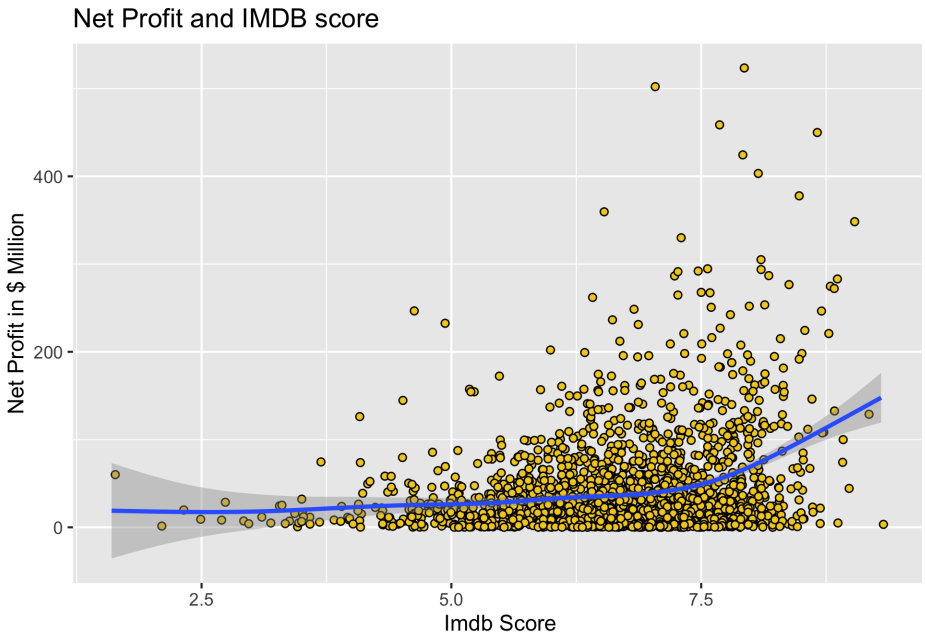


Impact of net profit on IMDB score.

Movies with a net profit over 200 million have a higher IMDB rating. Net profit is one of the strong indicators of a higher IMDB score. As the trend below shows higher net profits translates to a higher rating. It could be assumed that the viewership for movies with higher net profits was higher and thus, received a higher movie rating.

Technically, the movies with higher IMDB score should generate higher net profit. But this is not always true. There are many movies that have very good IMDB score but did not generate much profit. So, IMDB score cannot be a sole factor to consider the net profit. Sometimes, this score can be litte confusing as well.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Training the model

In order to get the most important variables, we are dividing the dataset into two. 80% of the data is divided as training dataset while the rest 20% is used for testing.

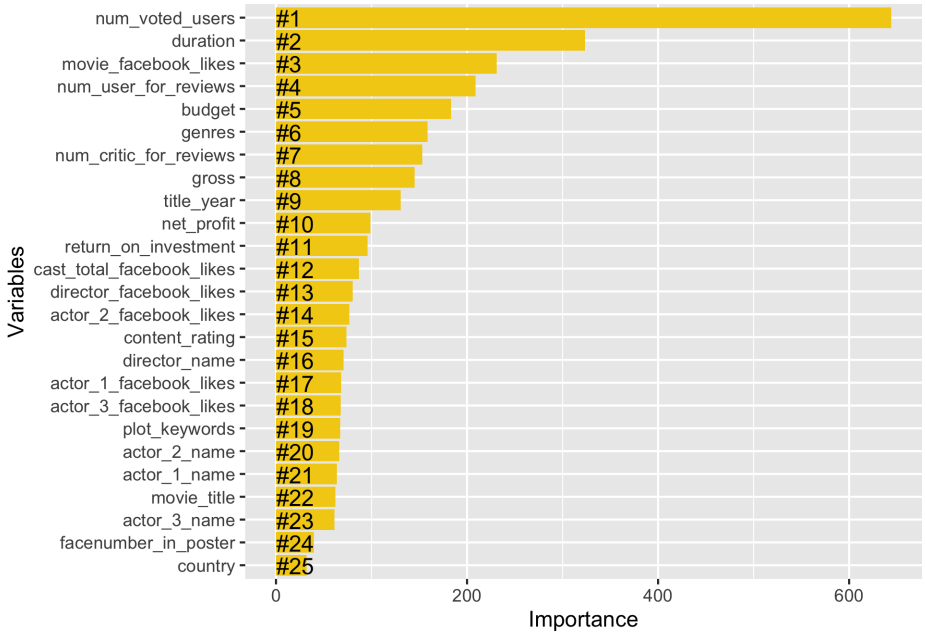
The linear model below shows that the number of voted users, number of critic reviews and duration has the most impact on the imdb score. The R-squared 0f 0.27 is extremely low which suggests the relationship between these variables is not linear even though they have a strong impact on the imdb score.

```
##
## Call:
## lm(formula = imdb_score ~ duration + num_voted_users + num_critic_for_reviews +
##     movie_facebook_likes, data = imdb_train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.888 -0.512  0.092  0.638  2.503
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.93e+00   8.44e-02   58.41   <2e-16 ***
## duration                1.07e-02   7.70e-04   13.85   <2e-16 ***
## num_voted_users         2.43e-06   1.37e-07   17.71   <2e-16 ***
## num_critic_for_reviews 4.19e-04   1.94e-04    2.17     0.03 *
## movie_facebook_likes   1.87e-06   1.14e-06    1.64     0.10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9 on 3039 degrees of freedom
## Multiple R-squared:  0.281,  Adjusted R-squared:  0.28
## F-statistic:  297 on 4 and 3039 DF,  p-value: <2e-16
```

Random Forest to determine the variable that has the most impact on the imdb score

We will be creating a random forest to identify the most important variables. For this project, we are building a random forest with 100 trees using all the variables. The model will be conducted on the training dataset.

```
##      |      Out-of-bag   |
## Tree |      MSE  %Var(y) |
##   10 |   0.7269    64.95 |
##   20 |   0.5934    53.02 |
##   30 |   0.5599    50.03 |
##   40 |   0.5436    48.57 |
##   50 |   0.5354    47.84 |
##   60 |   0.5217    46.62 |
##   70 |   0.5175    46.24 |
##   80 |   0.5147    45.99 |
##   90 |   0.5144    45.96 |
##  100 |   0.5094    45.52 |
```



We can see that the most important variable is the number of voted users. The reason is quite obvious because the rating only generates when people vote or give reviews for the movies. There might be a good number of people who don't giving reviews. But, there is also a good number of people who vote for their favorite movies which is used to calculate the imdb score. Second most important factor is the duration of the movies. This is quite interesting because this is not something which is easily guessed. However, the logical rationale behind this could be that the long hours movies are generally high budgeted ones with a renowned cast. Therefore, the quality of the movies with longer duration are usually better. The next factor is the facebook likes. Even though this factor is a difficult predictor to assume, we can say that people likes something on facebook when they truly enjoy something. So, they might have a tendency to provide good imdb rating as well.

The importance of next three variables, budget, genres and number of user reviews, is very close. A high budgeted movie will typically have a tendency to get high imdb scores because they are usually created with

with a lot of hype and promotion. Genres also have an impact because some genres are more attractive to users than others. Typically, action and thriller movies are preferred to many viewers. For obvious reasons, number of user reviews are important. These users directly rate the movies on the imdb website. So, certainly this will be one of the most important factors.

Conclusion

Random Forest took into consideration all the variables from the dataset to understand their impact on the imdb score. Number of voted users is the most important variable for a high imdb score. Followed by duration and facebook likes received by the audience. It is surprising to see, actors and directors names were amoong the least impactful factors as one would think directors and actors bring in publicity leading to a higher viewership.