**Analyzing IMDb Movie Scores: A Strategy for Motion Picture Studio Differentiation**

Rishi Malhotra, Yuyang Liu, Nick Ohadi, Hao Zhang, Xuanqi Liu

**Executive Summary**

In this analysis, our goal is to assess IMDb movie scores with the hope of providing a viable differentiation strategy for Disney Production Studios based on creating great movies (those with high IMDb movie scores). We used the software R to build five models and to calculate their responses to variables: three Multiple Regression models, one Decision Tree model, and one Random Forest model. Our modelling and findings revealed valuable insight that the following components/attributes of a movie were most important to increasing a movie’s IMDb score: the duration, gross revenue, budget, popularity of the director, popularity of the cast, genre, the MPAA rating, and the number of faces placed on a promotional poster. Therefore, to produce movies with potentially higher IMDb scores we recommend that Disney increase budget and film length, pick directors and cast members with a large number of Facebook likes, and avoid producing G-rated movies. In addition, production should focus on movies in the animation and drama genres as they tend to generate higher IMDb scores, thus greater movies.

**Dataset & Objective of Analysis**

We utilized a dataset from Kaggle.com titled “IMDb 5000 Movie Dataset”. This dataset included 28 different variables for 5,043 movies spanning across a time of 100 years in 66 countries. In addition, there are 2,399 unique directors and thousands of actors/actresses as well as 4,906 different promotional movie posters. The variables include: "movie\_title", "color" (whether movie is filmed in b/w or color), "num\_critic\_for\_reviews" (how many critics reviewed the movie), "movie\_facebook\_likes", "duration", "director\_name", "director\_facebook\_likes", "actor\_3\_name", "actor\_3\_facebook\_likes", "actor\_2\_name", "actor\_2\_facebook\_likes", "actor\_1\_name", "actor\_1\_facebook\_likes", "gross" (how much revenue the movie generated), "genres", "num\_voted\_users" (how many IMDB users voted), "cast\_total\_facebook\_likes" (combined # Facebook likes for all actors/actresses in the movie), "facenumber\_in\_poster" (how many actors/actresses were featured on the promotional movie poster), "plot\_keywords" (keywords to describe movie plot), "movie\_imdb\_link", "num\_user\_for\_reviews", "language", "country", "content\_rating" (MPAA rating), "budget", "title\_year", "imdb\_score", "aspect\_ratio" (aspect movie was filmed).

Prior to performing any analysis or modelling, it was necessary to clean the dataset. Any information that was missing or 0 for a unique record was coded as ‘None’ i.e. budget information. Furthermore, we accounted for inflation by using past consumer price indices to calculate the CPI for 2016 (since 2017 information is not yet available). Additionally, we decided to focus only on the films from the U.S.

To guide our analysis with a managerial perspective, we developed four research questions to address: (1) Can we predict a movie’s IMDb rating using its quantitative attributes? (2) Which genres are likely to correlate with great movies? (3) Will the number of human faces in a movie poster correlate with the movie rating? we utilized standard multiple regression models as well as standard machine learning techniques known as “Decision Tree” or “Random Forest”. Using the nearest neighbor predictor methodology the tree can advise us as as to whether a movie will have a good score based on the inputs we provide.

**Methodology**

*Multiple Regression*

The first method we used to analyze the data was multiple regression. We aimed to build a regression model that would help us explore the relationship between a variety of explanatory variables and the response variable, which is the IMDb rating assigned to each movie. Because there are so many variables to consider, some were deemed unnecessary and were excluded from the variable selection process. This includes Color, Director and Actor names, number of critic reviews, number of user reviews, plot keywords, and individual actor Facebook likes. These variables either held low explanatory ability, were captured in other variables, or would require too many dummy variables (e.g. individual actor names).

In order to manage the remaining variables, we utilized Forward-Stepwise regression. We split the remaining variables into three groups (Exhibit A) and put each group through Forward-Stepwise in order to build three regression models. The first group is the first group of variables, the second group adds the genres, and the third includes all the variables (adds the content rating variables). The resulting variables selected for each model can be seen in Exhibit B.

The data was split into a training set and a testing set, with a ⅔ to ⅓ split. The models were trained with the two-thirds of the data, and their accuracy was tested against the the testing set. We used Mean Average Percent Error to calculate the error and to compare which models were most accurate. The errors and adjusted R-Squared of each model can be found in Exhibit C. The errors were all relatively low, sitting between 10% and 12%.

The coefficients and details of Models 1, 2, and 3 can be found in Exhibits D, E, and F, respectively.

*Decision Tree*

The second method we used is Decision Tree. Since our analysis aims at explaining some important factors’ influence on the greatness of a movie, we believe Decision Tree is a good method to use due to its explicit decision rules. The following is the decision tree model we explored with the strongest prediction as well as explanatory ability:

* Dependent Variables: Moive Greatness Level. Specifically, we rounded the imdb-score to an integer to capture the movie greatness levels by 10 classes. For example, the 2.3 imdb score of a movie will be transformed to the second degree of greatness. And the distribution of the greatness level is relatively right-skewed. Most of the movies fall in the 5, 6 and 7 degree. (Exhibit G)
* Predictors: All 28 variables mentioned above.
* Calibration: First ⅔ of the data.

Exhibit H shows the outcome of the model above. It is notable that we only have classification rules for 5, 6 and 7 degrees. We believe this is because the sample size of the extreme ratings like 2 and 9 are too small to generate the common characteristics among them. From the tree (Exhibit H), we can infer that duration, movie\_facebook\_like, adj\_budget and the genres Drama and Animation have a positive impact on the greatness (IMDb score) of a movie. We also observed that there is kind of a complementary relationship among these attributes. For example, in the right branch of the tree, even though the duration of the movie did not reach 118.5 minutes, as long as it is over 104.5 minutes and its Facebook likes are over 13,500, it will still reach a 7 rating.

In terms of the performance, we used the mean of the percentage error across the validation sample to produce the error of the prediction, which is 12.25%. We also produced the confusion matrix for training data and testing data (Exhibits I & J). From the matrix, we can deduct that the most popular error range is 1. The model also tends to underestimate the rating of movies.

*Random Forest Regression*

The third method we used is the Random Forest regression. Before fitting the model, we categorized the scores into grades A to E, setting intervals to be 2. After the categorizing, we found that 60% of the movies belong to Grade B (score 6 to 8), while about 5% of the movies are rated below 4 or over 8. The model was fitted to predict movie scores using the following variables:

* imdb\_grades: ratings 0 to 10 are divided into 5 levels of grades
* director\_facebook\_likes: the number of likes on the director’s Facebook page
* cast\_total\_facebook\_likes: the number of Facebook likes for all cast members
* duration: the length of movies
* title\_year: release year
* facenumber\_in\_poster: the number of faces on the IMDb main poster
* adj\_budg: the measure of budget that takes into account the time period's inflation rate
* genres (21 genres): 21 genres categorized by IMDb

Since running a random forest regression does not need cross-validation or a separate test dataset to estimate an unbiased test dataset error, we used the entire dataset which includes 1494 observations, to grow up to 10,458 trees to fit the random forest. The number of variables tried at each split of the decision tree is 5. The out-of-bag error rate is 29.38%, and the class error for A to E grades are 97.44%, 14.13%, 46.33%, 89.19% and 100.00% (Exhibit K). The error rate for the random forest regression is about 18% higher than the one generated by the multiple regression model, but random forest is still the optimal approach to predict movie grades utilizing the information we have before movie production.

**Key Findings**

The decision tree, random forest, and the various regression models revealed insightful findings. Namely we discovered which significant variables (those with p-value i.e. <0.005) are responsible for decreasing or increasing a movie’s rating. The variables that will decrease a particular movie’s IMDb score/rating are: "facenumber\_in\_poster" (as seen in Regression Model 1, Exhibit D), the number of faces that are included in the promotional materials and posters (the -ve coefficient indicates that more faces, the lower the score), and certain "content\_rating" variables (as seen in Regression Model 3, Exhibit F); a movie receiving a G-rating by the MPAA will have a low IMDb score. If any of the following variables are increased: "duration", "gross", "budget", "director\_facebook\_likes", or "cast\_total\_facebook\_likes" then the IMDb score of the movie will increase, based on +ve coefficients (see Regression Models, Exhibits D, E, and F).

From the fitted random forest model, the importance of each variable is shown in the graphs below (Exhibit L). Duration has the largest impact on movie grade among all the variables, followed by the number of likes on the director’s Facebook page and the adjusted budget of movies. The popularity of the cast and release year also contribute to the movie grade to a certain extent. However, number of faces in a poster does not have sufficient impact on the grades. According to the random forest model, none of the genres are good predictors. This finding conflicts with the results generated by the other models, since we took gross revenue out of our dependent variables selection and applied a different approach. By doing so, we can better predict the movie grades before the release of movies. Also different from the multiple linear regression models, the "adj\_budg" is considered as a critical variable, even a little bit more important than the "cast\_total\_facebook\_likes".

The second multiple regression model (Exhibit E) provided insight as to which genres are likely correlated with great movies. We selected genres that were statistically significant (with a p-value <0.005) through the forward stepwise-regression phase (Exhibit A) in model two. We found that Animation, Drama, Comedy, and Horror are all significant, whereas Fantasy was statistically insignificant. After interpreting regression model three (Exhibit F), we found similar findings to model two except that the genre Fantasy was marginally significant and Music was not significant. Therefore, the results indicate that the genres Animation and Drama are likely to have higher IMDb ratings (or increase IMDb score due to +ve coefficient) and that the genres Horror and Comedy are likely to have lower IMDb ratings (or decrease IMDB score due to -ve coefficient). To further confirm these results, our decision tree model (Exhibit H) also reflects the effects of Animation and Drama, that these two genres have a positive effect on IMDb score. Thus we can conclude great movies are associated with the genres Animation and Drama.

Lastly, we concluded that the fewer the number of faces/actors on a movie poster, the greater the IMDb score (Exhibit N). Our conclusions for the effect of the number of faces in a movie poster on the IMDb score were based only on regression Model 1 (Exhibit D), since it had the lowest error rate. We found that the coefficient of the variable is -0.0240, indicating that it has a negative effect on the dependent variable, but that the p-value is only marginally significant, 0.0502. There is likely a “sweet spot” for the number of faces to include in a movie poster, but the model’s coefficient does not reflect this as it would imply zero faces is the best.

**Recommendations**

Our key findings demonstrate that there are a few areas that Disney management and the production studio can focus on in order to make great movies (movies with high IMDb scores). First, we recommend the studio to focus its efforts on producing Animation and Drama genre movies in the future. By doing so the studio can earn high ratings for its movies and differentiate from other studios by making great movies. In addition, the motion picture studio and marketing teams can focus on cross-selling and promoting movies in either of these genres as it will likely boost appeal and rating for both movies due to their association. The Horror and Comedy genres should generally be avoided if the goal of the studio and producers is to make great movies. However, Horror and Comedy should not be overlooked completely, since they can potentially be profitable genres that appeal to the consumer market especially attractive segments.

Second, we recommend the studio take control of several key variables in the production process of movies that will result in achieving higher IMDb scores for its movies. The studio can directly control the following attributes of a movie before production: the duration of the film, the director(s), the various cast members, the content-rating (in association with the MPAA), as well as how much budget to allocate to the movie. Specifically, by increasing the film’s length, picking directors with a high number of Facebook likes, as well as cast members with a high number of Facebook likes, ensuring a non-G rating, and allocating higher budgets to movies, the studio can produce a great movie.

Last but not least, in order to guide creative advertising and promotional teams at the studio, we advise creating promotional movie posters with a theoretical sweet spot in mind for the number of faces/actors to include. Since there is a direct correlation between the number of faces included on a movie poster and how great of an IMDb score the movie receives, we recommend creating additional marketing materials (flyers, billboards) to experiment with the different number of actors/actresses faces to include on posters. We recommend that further research be done on the design of posters before the Disney Studio commits to reducing the number of faces in their posters. There are numerous attributes to a movie poster’s design that can affect a movie’s rating, and that can interact with the number of faces in the poster. These includes: color scheme, composition, which actors’ faces are displayed, size of faces, and so forth. For example, *Star Trek II* and *500 Days of Summer* both received an IMDb rating of 7.7. However, *Star Trek II*’s movie poster contains only three large faces of key actors, whereas *500 Days of Summer* includes 43 faces - one face of the male lead, and numerous of the female lead’s face. The way faces are utilized in both of these posters are vastly different - see Exhibit M for the comparison between these two posters.

**Appendices**

Exhibit A: Variable Groups for Forward-Stepwise

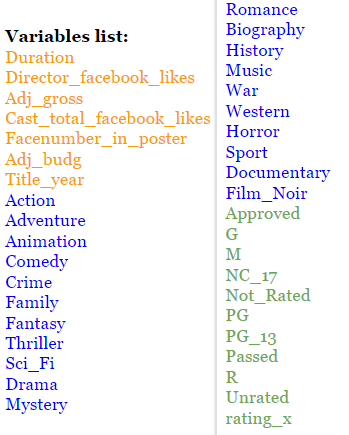


Exhibit B: Selected Variables for Each Regression Model

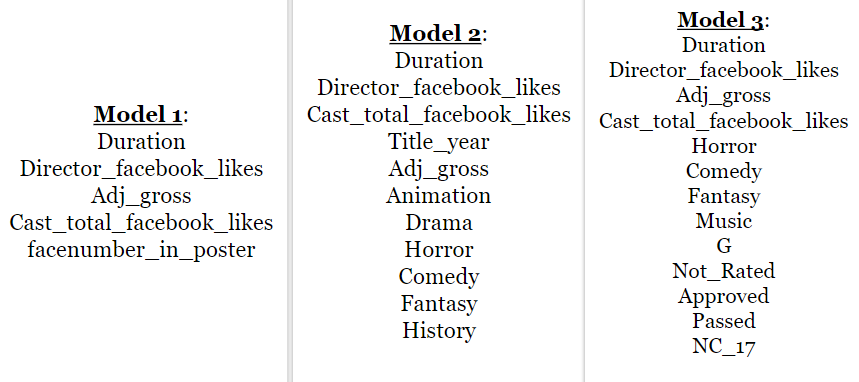


Exhibit C: Errors and R-Squared of Regression Models

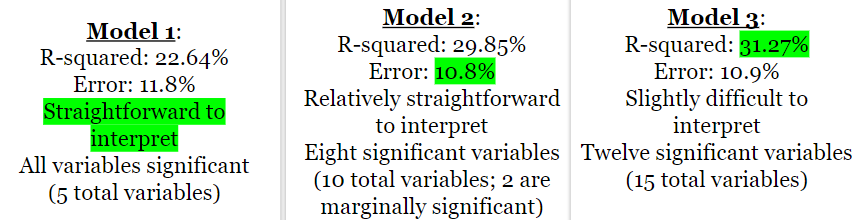


Exhibit D: Regression Model 1

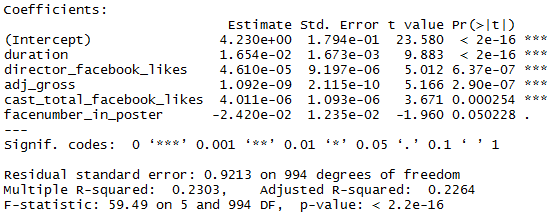


Exhibit E: Regression Model 2

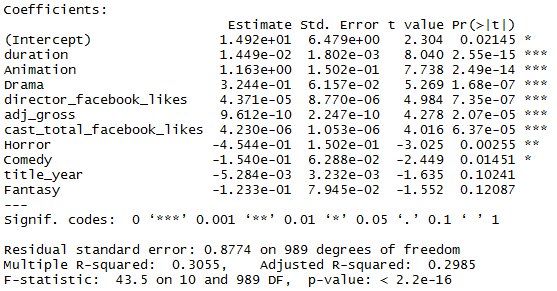


Exhibit F: Regression Model 3

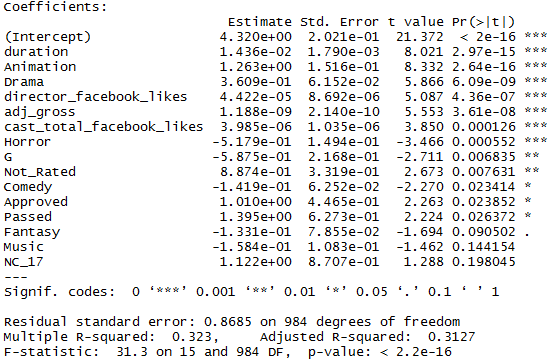


Exhibit G: Greatness Level Distribution

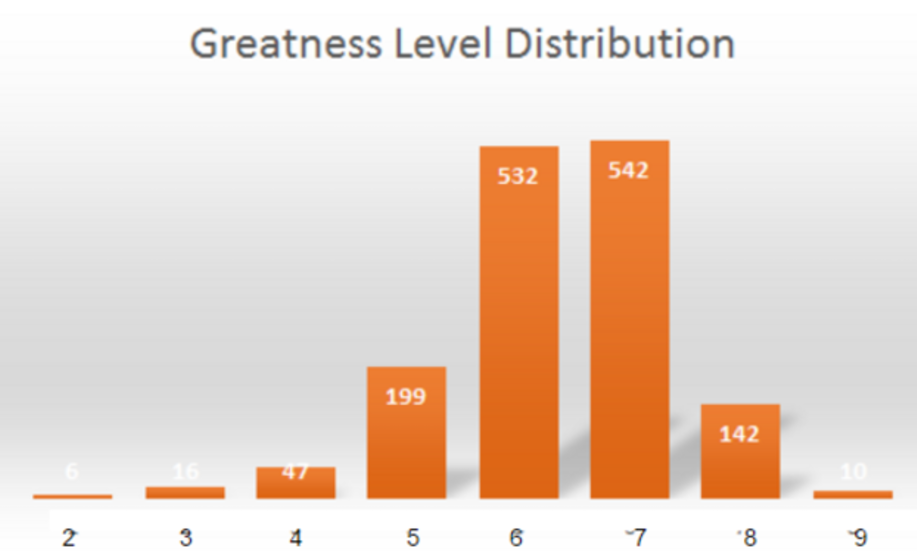


Exhibit H: Decision Tree

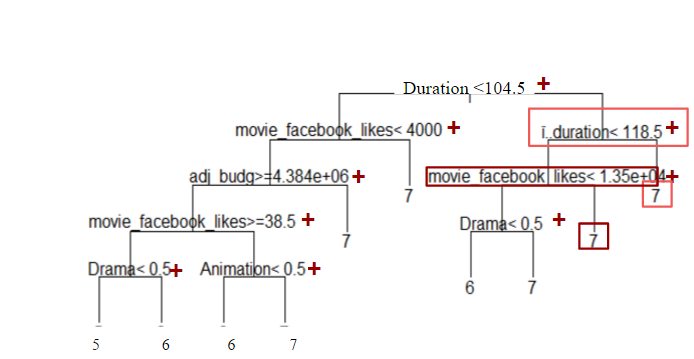


Exhibit I: Confusion Matrix of Training Data

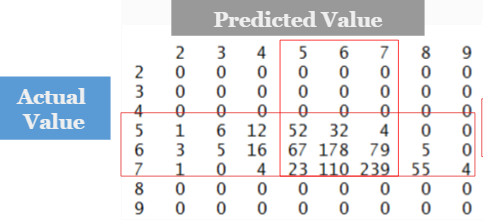


Exhibit J: Confusion Matrix of Testing Data

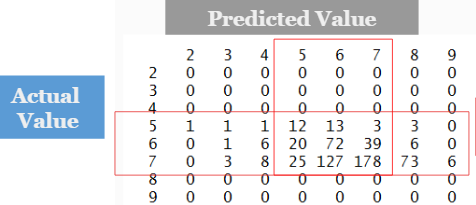


Exhibit K: Confusion matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | E | D | C | B | A | class. error |
| E | 0 | 0 | 0 | 1 | 0 | 1.000 |
| D | 0 | 4 | 23 | 10 | 0 | 0.892 |
| C | 0 | 0 | 278 | 240 | 0 | 0.463 |
| B | 0 | 0 | 125 | 772 | 2 | 0.141 |
| A | 0 | 0 | 0 | 38 | 1 | 0.974 |

Exhibit L: GINI

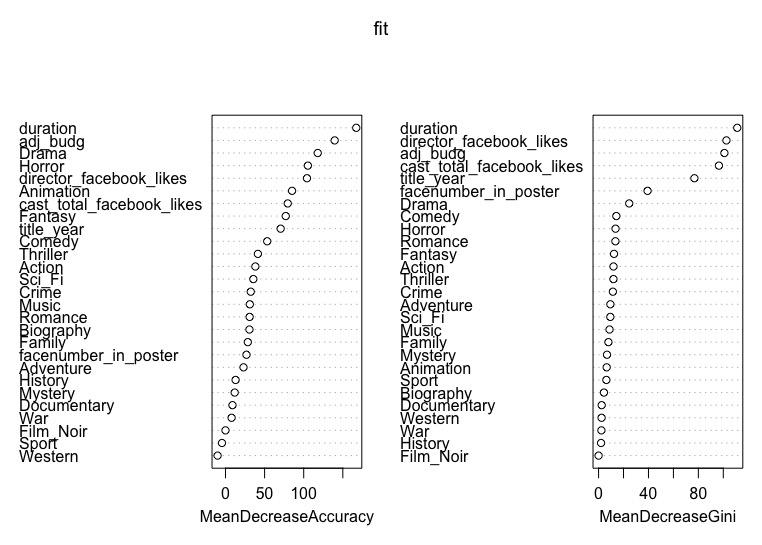


Exhibit M: Movie Poster Comparison



Exhibit N: Number of faces in Poster vs. IMDb score

