**Prediction Model for IMDB Movie Ratings to Enhance Robustness against Biased Vote Stuffing**

Chaitnay Singh, Sanchit Khattar, Parijat Rai, Joseph Tsai

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

[singh488@purdue.edu](mailto:singh488@purdue.edu), [rai21@purdue.edu](mailto:rai21@purdue.edu), [skhattar@purdue.edu](mailto:skhattar@purdue.edu), [tsai103@purdue.edu](mailto:tsai103@purdue.edu)

**Abstract**

Accurate movie ratings are crucial to IMDb because users visit their website in search for this genuine metric. Hence, without accurate ratings, IMDb could not attract sufficient users to sustain its website’s advertising value. Currently, high degrees of vote stuffing has emerged on the website, and the IMDb rating systems is confronted with the threat of losing its accuracy. Our team has created an regression model with an adjusted R square of 51% that looks into more diverse attributes to provide robust ratings against vote stuffing. Our model serves as an initial step in the exploration of creating a more balanced rating system.

**Keywords:** IMDb, Movie Ratings, Regression Model, Predictive Analytics.

**Business Problem:**

With more active users and parties of interest streaming on to the IMDb platform, movie ratings are becoming harder to balance because the scale of “vote stuffing” has gotten greater. For example, a poor quality movie can be rated as a great movie due to more users who intentionally generate multiple high ratings. The current method that IMDb adopts is weighted average, certain users account for more when the ratings are aggregated. The weighted average method was applied in hope to filter out part of the voters who intentionally give biased ratings. However, new methods of balancing the movies ratings are required in order to maintain the fair and welcoming environment IMDb tailored for movie lovers.

**Analytics Problem:**

Leveraging the information of previous movies and their ratings, our team has discovered certain characteristics that exist in popular movies and certain characteristics that are more related to poor rated movies. In order to provide a reference score to strengthen the current rating system, the challenge remains in how we utilize these diversified characteristics to form an accurate prediction model that is robust against biased votes/ratings.

**Data:**

The dataset we used is the IMDb 5000 Dataset which is scraped from the IMDb website. The dataset contains general information of 5043 movies with release dates from 1916 to 2016. The main variables that we analyzed are as follows:

Genre- The style and elements of the movie, such as action/ animation/comedy/thriller.

Budget- The net budget of producing the movie.

Number of critics- The number reviews made by critics which the movie has received.

Duration- The length of the movie.

Number of users review- The numer of reviews made by normal users.

Number of voted users- The number of users who voted for the movie.

Number of faces in posters- The number of faces that appeared in the movie’s poster

Movie facebook likes- The number of followers the movie has on facebook.

Gross- The amount of revenue the movie generated in theaters.

Director Facebook Likes- The number of followers the director has on facebook.

Actor Facebook Likes- The number of followers the actor has on facebook.

Since the movie information dates from 1916, we incorporated an inflation rate dataset in order to standardize the currency to the current USD value.

**Methodology Selection:**

In the beginning, our team explored two models to help facilitate accurate ratings: the regression model and the decision tree model. Both models were equally accurate in the way they provided similar root mean square error(RMSE). However, we chose the regression model in the end because it can more clearly reflect the correlation of each attribute with the movie ratings and is also easy to comprehend.

**Model Building:**

Dataset Preparation: First, our team standardized the currency value across the timeline to the current monetary value(USD) by multiplying the values by inflation factor. Second, we created dummy variables for the movie genres. Third, we randomly drew 75% of our dataset as our training dataset, and the remaining 25% of our data served as testing dataset.

Exploratory Analysis: After the dataset preparation, we drew histograms to visualize the distribution of movie ratings. Furthermore, we examined scatter plots to have an initial understanding of the correlations between rating and each attribute.

Model Refinement

Initially, our regression model consist of most of the attributes/variables we have, reaching an adjusted r square around 46%. After that, we dropped the variables that are less significant, and continuously tested on the testing dataset to finalize the regression model. Ultimately, our final movie rating prediction model reached a adjusted r square of 51%.

**Functionality:**

With variables such as genre, budget, and number of facebook likes and etc., our model would be able to generate a prediction of the movie rating with an adjusted r square of 51%. With the diverse attributes we incorporated, the rating score that our model produced can serve as a reference to balance the existing ratings that are solely generated from users and critics.

**GUI Design and Functionality:**

Our team designed the UI of our app in accordance with the style of the IMDb website. The theme of our model’s UI design is “theater vibe”. We presented this models with lively colors and fonts that are often seen in movie theaters. On the UI surface, we employed sliders, radio buttons, select inputs and numeric inputs for suitable attributes, allowing users to enjoy the simplicity with the sophistication of data collection.



                                                    Full screenshot of the UI where users can input relevant information and predict movie ratings.  
**Conclusions:**   
Our model is an initial project that opens new possibilities for future rating systems. We were able to achieve an adj R square of 51% based on diverse attributes. In the future we would like to use more sophisticated models like random forests to achieve better accuracy. Also attributes like Facebook Likes are currently biased as new movies tend to get a higher number of likes since some of the old movies did not have social media during their time and so the latest trending movies tend to get higher likes. The dataset also has scope for text/sentiment analysis and we would like to incorporate this in the future. Our model currently has high heteroskedasticity which is something we would like to control in the latter versions. In the future, models like this can be further developed to help provide better ratings that are robust against the noise and extreme values within the data records.

**References:**

Why You Can’t Trust User Ratings from IMDB: A Case Study by Derek Jacobs <https://plotandtheme.com/2017/01/19/why-you-cant-trust-user-ratings-from-imdb-a-case-study/>

The vote average for film or TV show "X" is clearly wrong! Why are you displaying another rating? by IMDb <http://www.imdb.com/help/show_leaf?votes>

IMDb Votes/Ratings Top Frequently Asked Questions by IMDb <http://www.imdb.com/help/show_leaf?votestopfaq>

TMDB 5000 Movie Dataset on Kaggle <https://www.kaggle.com/tmdb/tmdb-movie-metadata>

Predicting IMDb ratings using numerical attributes by NilayShrivastava <https://www.kaggle.com/nsnilay/predicting-imdb-ratings-using-numerical-attributes/code>