Movie Classification & Analytics

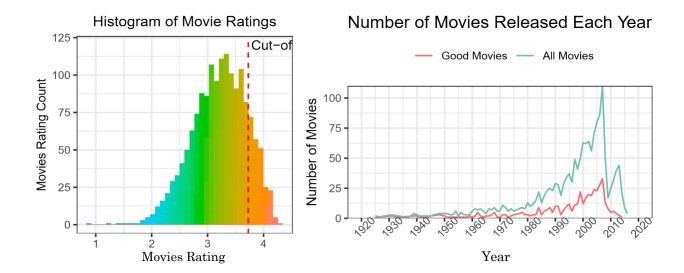
Authors: Yin, Karolina, Emma, Rachana, Nupur, Sarah

2023-02-18

Raw Data Summary and Cleaning

The Full MovieLens Dataset has 26 million ratings and 750,000 tag applications from 270,000 users on all the 45,000 movies. Before training predictive models, we cleaned the raw data by dropping irrelevant columns or columns with over 50% missing values, eliminating visually extreme values and created new metrics/features (weighted rating, ROI, profitability, etc.)

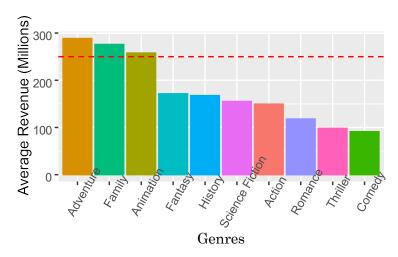
Data Analytics



Understanding Movies Rating and History

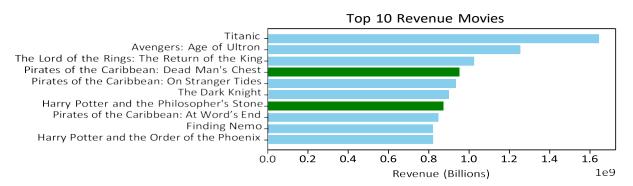
Based on IMDB's weighted rating (wr) formula, we created a new fairer feature "weighted_rating" which takes number and average of voter ratings into consideration during calculation. Having observed below histogram, it is safe to argue that vast majority of movies rating are above 3.0. Therefore, in order to set apart "good" movies, we need to establish a criteria determining ultimate result, which is "if weighted_rating > cut-off 3.54, and vice versa.) Specifically, the cut-off is derived from 85th quantile of weighted rating, meaning if a movie's weighted rating > 3.54, then it is better than 85% all movies, thus rendering itself as "good movies". To identify factors that may influence a movie's rating, we considered various explanatory variables, such as the cast, budget, production company, release date, number of competing movies released at the same time, and movie length (runtime)

Top 10 Genres by Average Revenue



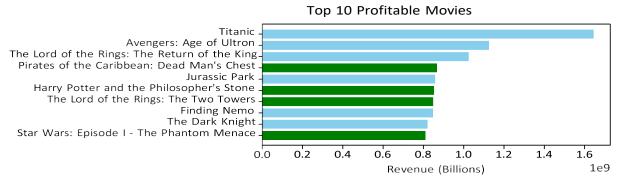
Findings on revenue vs genre:

Adventure, Family, Animation genres generally generate the most amount of revenue. It is interesting to notice that Action, Adventure, and Drama genres also score top 10% in terms of the length of the movie, which can result in higher popularity due to length preferences in comparison with other genres. This probably resulted in increase in revenue.



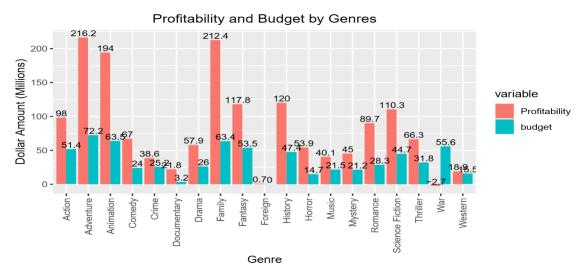
Findings on revenue vs movies:

For Top Revenue Movies it turned out that Titanic and Avengers scored highest for the revenue but 2 movie Pirates of the Caribbean Dead's Man's Chest, Harry Potter and the Philosophere's Stone had highest ratings above 3.54 considering they are a good movie. This can be attributed to the cause of different audience demographics with varying preferences but interestingly Pirates of the Caribbean might have highest rating due to its genre adventure and its length which can be related to the insight in 2.2.1



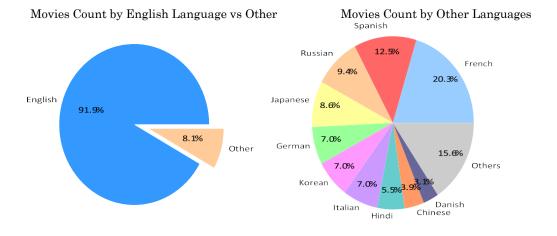
Findings on profitability vs movies:

Top Profitable Movies that present four good movies with rating > 3.54 are marked in green. These four movies fall under the genre of Adventure. Interestingly, Adventure genre again appears to be on top in terms of Profitability supporting insights 2.2.1 and 2.3.1.



Findings on Profitability/budget vs genres:

Adventure, Family, and Animation genres have the highest profitability. Interestingly, Adventure genre again shows that not just in tops in revenue but by spending small budget profit and revenue can be easily maximized in this genre. This again supports all the above insights.





Findings on Word Cloud:

Turns out that city, name and family matters relate to higher rating. Overall, Adventure and Family genre movies rank high in revenue, profitability, length and rating.

Model:

Logistic model was used to determine which variables have a significant impact on a movie's rating and to draw meaningful insights from the data. Six logistic models had been trained, and the sixth one was selected as final model. In comparison with other models, Model 6 has the lowest Akaike Information Criterion (AIC) which is a measure of the relative quality of statistical models. A lower AIC value indicates a better fit of the model to the data. This is striking a good balance between simplicity and complexity. The selected model's (lowest AIC) deviance residuals indicates that the model is a reasonable fit to the data, and among all factors that contributing to the probability of predicting good movies, we can qualitatively calculate the marginal effect of all significant predictor variables. For example, the estimate for the drama predictor variable is -0.4369. This means that for every one-unit increase in the drama predictor variable, the log-odds of the outcome decreases by 0.4369, while holding all other predictor variables constant. In other words, an increase in the drama variable makes the outcome less likely.

The model provides important insights on which features are important for predicting the success. Based on the model, we can see that the variable rating has the highest positive impact on the success of a movie, which means that movies with a higher number of ratings (indicating higher popularity) are more likely to be successful.

Conclusion:

In conclusion and comparing our findings, Adventure and Family genre movies tend to be highly profitable and have a strong market demand. They also have high rating and if we look at model 6, rating feature is important for predicting movie success. They have a wide appeal and can attract a broad audience. These movies tend to resonate emotionally with viewers and the emotional connection can lead to a more engaged audience, again making it profitable. After analysis we found that these genres have consistently performed well in terms of box office revenue and audience ratings. Additionally, our analysis revealed that longer movies tend to perform better at the box office. Movies that are around two hours or more tend to have higher box office revenue and audience ratings than those that are shorter in length. It is due to immersive experience by viewers. Given these findings, we recommend that the production team focus on developing movies within the adventure, family and action genres, and consider creating longer films to maximize profitability.