

Box Office Busts

Jeremy L. Smoljan

University of Nevada, Las Vegas

IS 489 1001 - Advanced Business Analytics

Professor Sutirtha Chatterjee

12 December 2025

Abstract

Utilizing Microsoft Excel, RStudio, and Rapidminer to discover major factors affecting box office revenue and successful outcomes for movies. This analysis showcases an overall view of the instability in predicting an accurate gross revenue in the box office while highlighting the importance of specified variables in predicting a ‘better’ performing movie.

Box Office Busts

Topic 1: ‘How can you predict which movies will be a hit or a bust?’ This analysis discovers the major factors affecting how well a movie will generate revenue in the box offices. In the question being addressed, this analysis is an in-depth study of ‘why’ movies succeed and fail in the manner of how much revenue was generated in correlation to specified variables, as opposed to the other topic wherein a marketing budget is determined in ‘how’ to be used. So, in a bid to better understand and develop functional skills in analysis, this topic was chosen. In this ‘why’ versus ‘how’ scenario, the former was selected to improve hard skills.

Development

Analysis Techniques

Due to the substantial amount of independent variables and missing data, random forest models and gradient boosted trees were utilized to cut down on overfitting and to assist in regression accuracy. Random forest analysis’ and Gradient Boosted Tree models also helped in directly highlighting which features are important when determining ways to increase box office revenue, or in this case, ‘which movies will be a hit or a bust?’

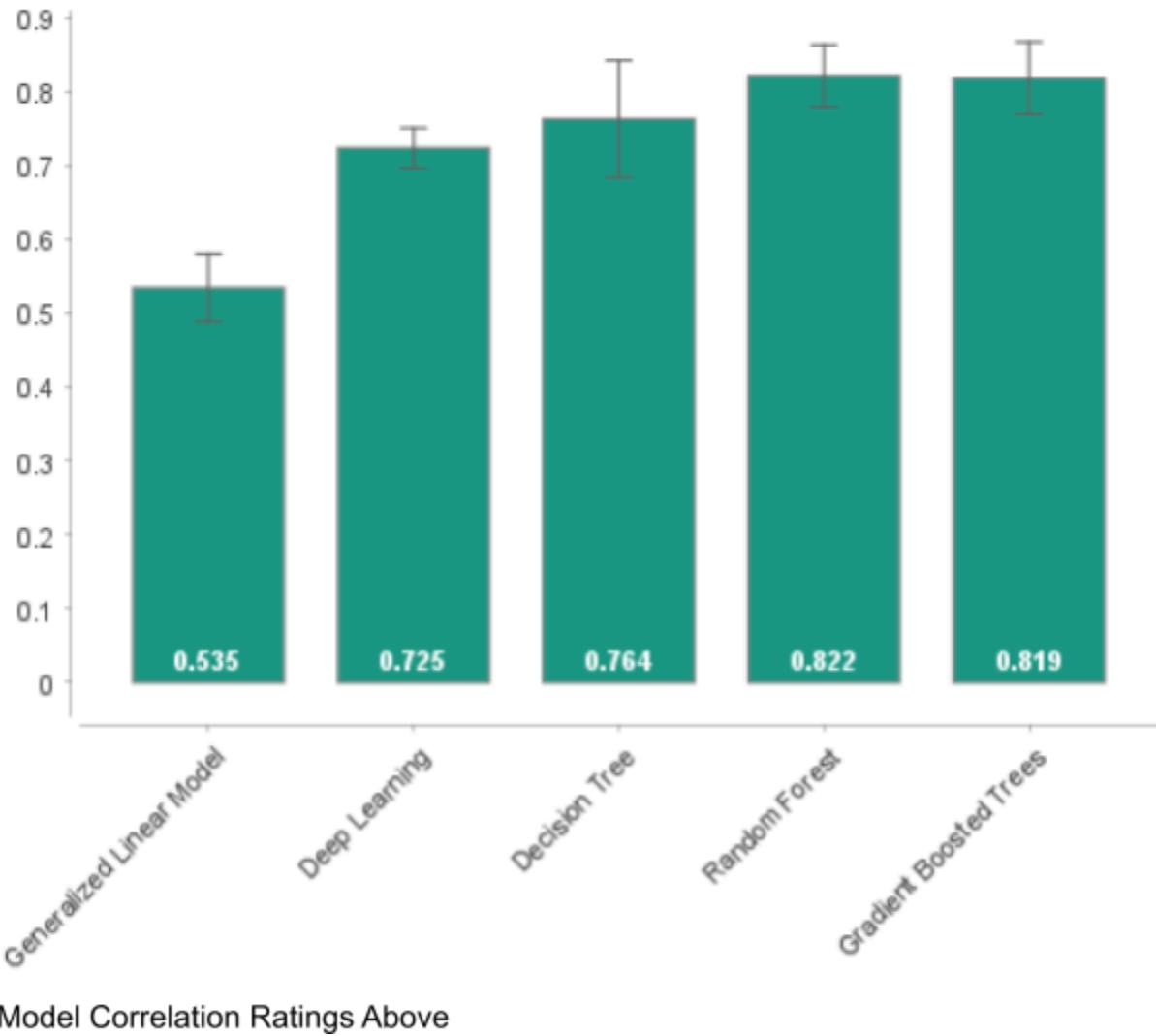
Tools Used

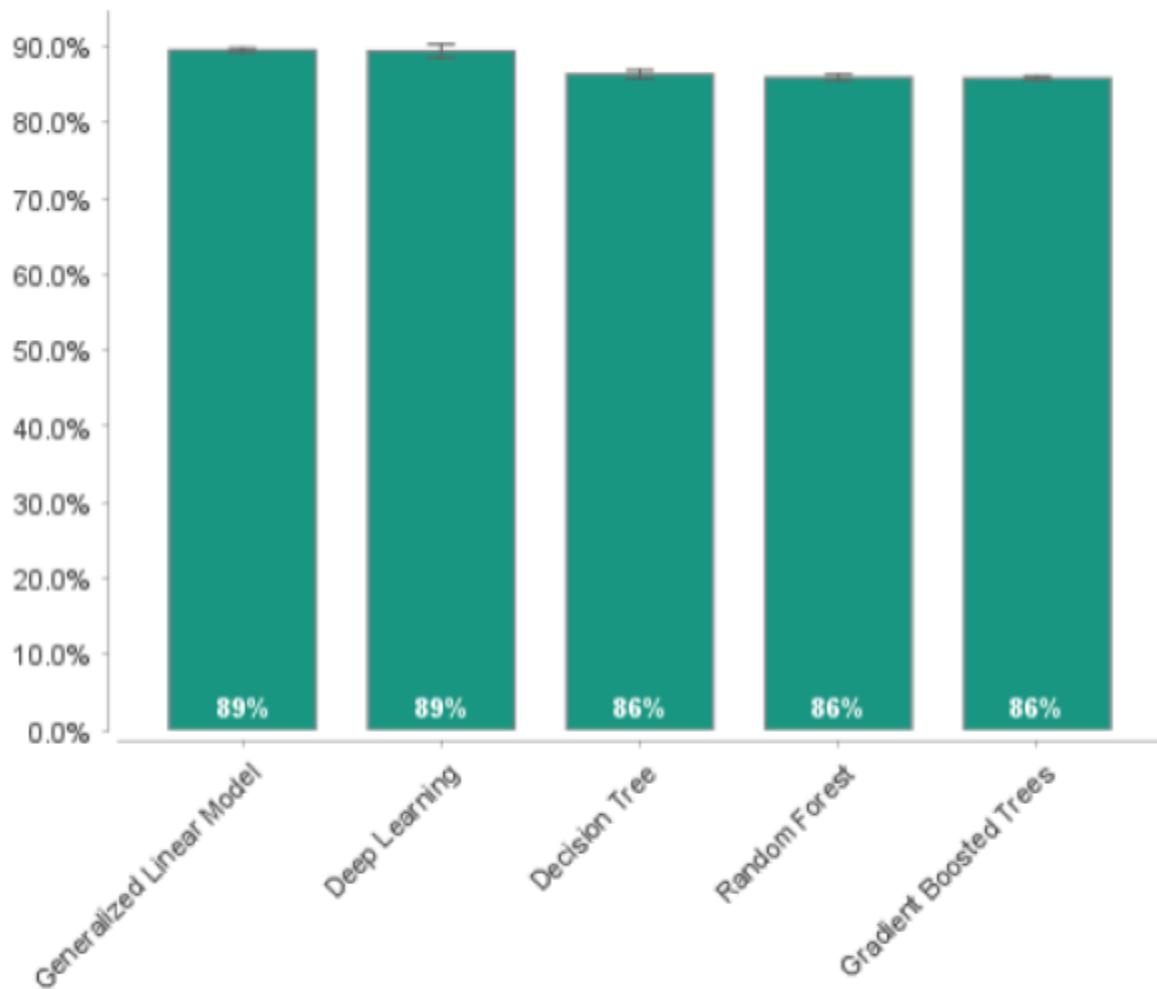
Rstudio was utilized to clean and refine the data alongside Microsoft Excel. First, the data format was cleaned by converting GBP and EU currencies to USD. Then the date format was changed to yyyy-mm-dd to work with Rstudio. Finally, the columns were cleaned of mixed variables, including mostly ‘N/A’ entries located throughout the numerical columns. Then Rstudio was used to create a season column to separate the ‘release_date’ entries into an easier access format for Rapidminer revolving around the four seasons, along with creating three new

columns out of ‘keyword’ for ‘sequels’, ‘remakes’, and ‘tentpoles’ utilizing binary 0/1 formats.

After cleaning and transforming the data, Rapidminer was used for analyzing and modeling.

Analysis - Graphs, Pictures, Tables, Figures





Model Relative Error Performance Above

Weights by Correlation

Attribute	Weight
rating	0.377
sequel	0.334
imdb_rating	0.129
remake	0.125
season	0.073
metacritic	0.064
release_month	0.047
Budget	0.027

Combined Overall Weights by Correlation Above

Gradient Boosted Trees - Weights

Attribute	Weight
Budget	0.415
metacritic	0.249
rating	0.083
season	0.031
remake	0.024
imdb_rating	0.004
release_month	0.000
sequel	0.000
tentpole	0

Independent Variable Weights within Gradient Boosted Trees model Above

Random Forest - Weights

Attribute	Weight
Budget	0.614
metacritic	0.320
remake	0.104
rating	0.033
season	0.031
imdb_rating	0.019
release_month	0.003
sequel	0.000
tentpole	0

Independent Variable Weights within Random Forest model Above

Correlations

Overall Matrix Correlations Above

root_mean_squared_error

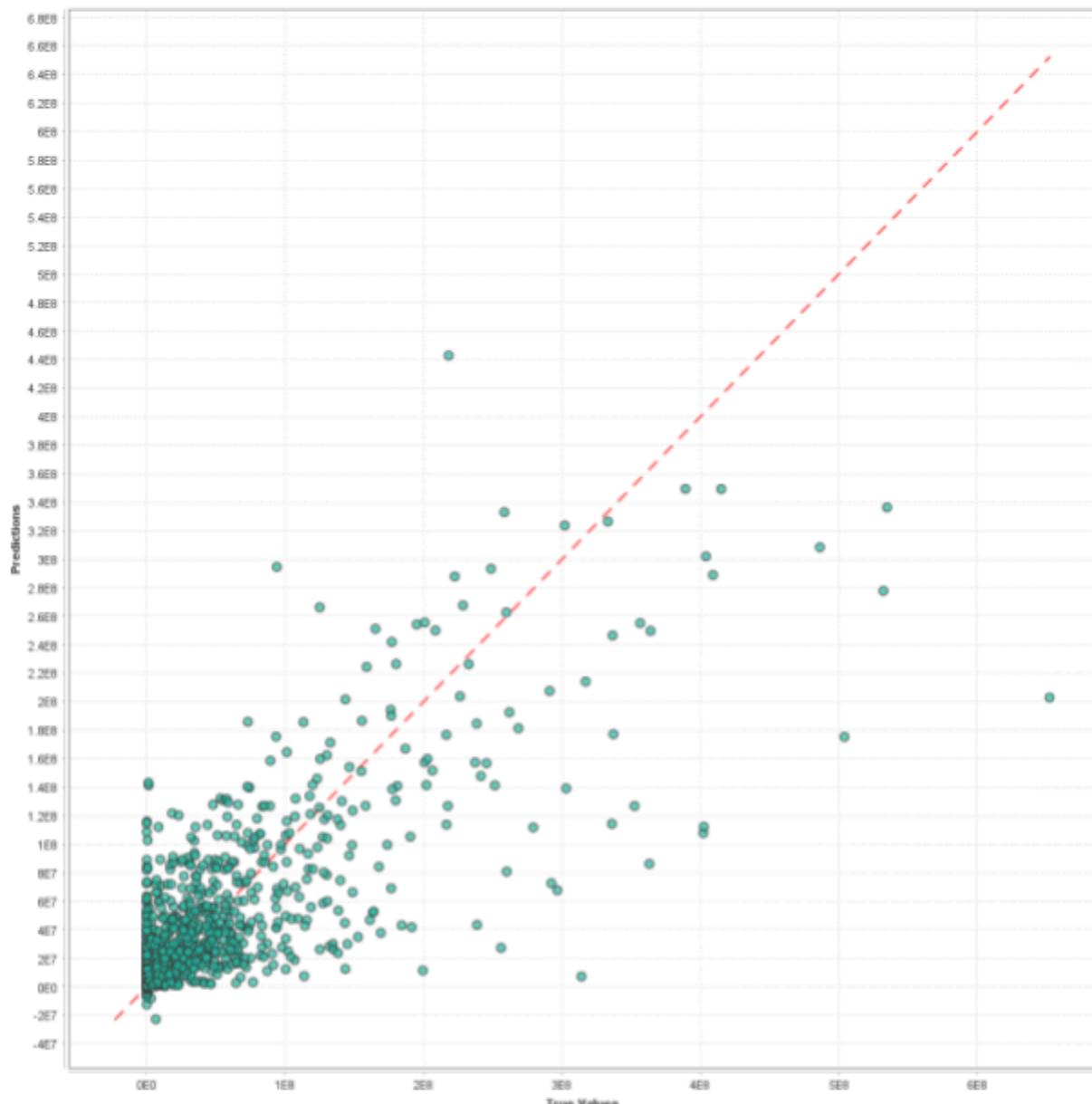
root_mean_squared_error: 26729406.044 +/- 2292388.804 (micro average: 26808582.446 +/- 0.000)

correlation

correlation: 0.819 +/- 0.049 (micro average: 0.818)

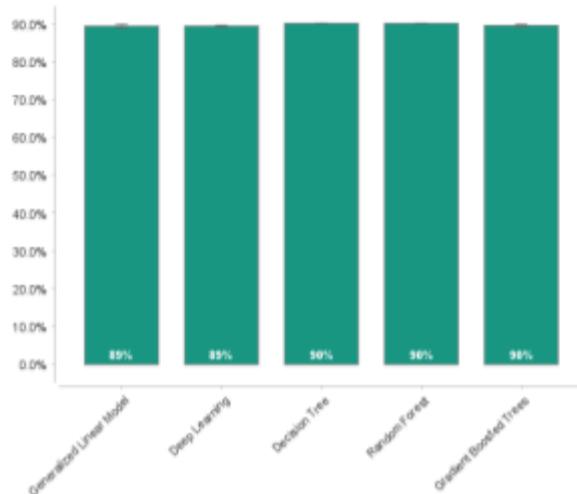
RMSE and Correlation within Gradient Boosted Tree Model Above

Gradient Boosted Trees - Predictions Chart

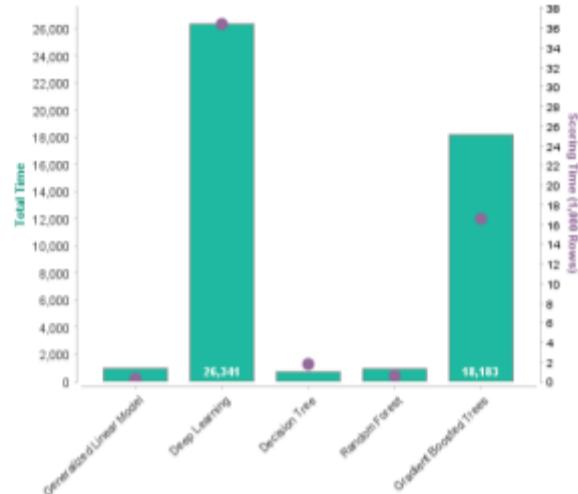


Number of Models: 7

Relative Error



Runtimes (ms)



Relative Error

Model	Relative Error	Standard Deviation	Gains	Total Time	Training
Generalized Linear Model	89.4%	± 0.4%	?	968 ms	9 ms
Deep Learning	89.4%	± 0.2%	?	26 s	995 ms
Decision Tree	90.1%	± 0.0%	?	712 ms	~0 ms
Random Forest	90.1%	± 0.0%	?	931 ms	~0 ms
Gradient Boosted Trees	89.6%	± 0.3%	?	18 s	40 ms

Genre model relative error values Above

Correlations

Genre General correlations matrix between provided models, Left

Attributes	Box.Offi...
Box.Office.Gross	1
genre = Action	-0.028
genre = Action, Adventure	0.059
genre = Action, Adventure, Biography	-0.002
genre = Action, Adventure, Comedy	0.040
genre = Action, Adventure, Crime	0.050
genre = Action, Adventure, Drama	0.065
genre = Action, Adventure, Family	0.073
genre = Action, Adventure, Fantasy	0.202
genre = Action, Adventure, History	-0.006
genre = Action, Adventure, Horror	0.024
genre = Action, Adventure, Mystery	0.087
genre = Action, Adventure, Romance	0.002
genre = Action, Adventure, Sci-Fi	0.259
genre = Action, Adventure, Thriller	0.054
genre = Action, Adventure, Western	0.012
genre = Action, Biography, Crime	-0.010
genre = Action, Biography, Drama	0.065
genre = Action, Biography, History	?
genre = Action, Comedy	0.018
genre = Action, Comedy, Crime	0.032
genre = Action, Comedy, Drama	-0.021
genre = Action, Comedy, Family	0.001
genre = Action, Comedy, Fantasy	0.015
genre = Action, Comedy, History	-0.004
genre = Action, Comedy, Horror	-0.006
genre = Action, Comedy, Music	0.003
genre = Action, Comedy, Musical	-0.004

Gradient Boosted Trees - Predictions

Row No.	Box.Office.Gross	prediction(Box.Office.Gross)	genre
1466	113203870	131685325.311	Action, Adventure, Mystery
1989	402453882	131685325.311	Adventure, Sci-Fi, Thriller
2627	89021735	131685325.311	Action, Adventure, Mystery
2783	0	131685325.311	Adventure, Sci-Fi, Thriller
2898	66002193	131685325.311	Action, Adventure, Mystery
523	124987023	123249466.593	Adventure, Drama, Fantasy
536	292298923	123249466.593	Adventure, Drama, Fantasy
537	296623634	123249466.593	Adventure, Drama, Fantasy
789	2184640	123249466.593	Adventure, Drama, Fantasy
526	184208848	121922920.561	Adventure, Drama, Western
2378	0	121922920.561	Adventure, Drama, Western
870	228433663	84160296.030	Adventure, Drama, Sci-Fi
76	0	80511462.093	Action, Adventure, Sci-Fi
267	632177324	80511462.093	Action, Adventure, Sci-Fi
381	0	80511462.093	Action, Adventure, Sci-Fi
538	337136885	80511462.093	Action, Adventure, Sci-Fi
761	6820	80511462.093	Action, Adventure, Sci-Fi
1470	333176600	80511462.093	Action, Adventure, Sci-Fi
1705	132556852	80511462.093	Action, Adventure, Sci-Fi
1708	158848340	80511462.093	Action, Adventure, Sci-Fi
1710	176654505	80511462.093	Action, Adventure, Sci-Fi
1719	314057748	80511462.093	Action, Adventure, Sci-Fi
1825	0	80511462.093	Action, Adventure, Sci-Fi
1843	0	80511462.093	Action, Adventure, Sci-Fi
1953	57761012	80511462.093	Action, Adventure, Sci-Fi
1965	103144286	80511462.093	Action, Adventure, Sci-Fi
1967	116601172	80511462.093	Action, Adventure, Sci-Fi
1972	148408305	80511462.093	Action, Adventure, Sci-Fi
1983	245439076	80511462.093	Action, Adventure, Sci-Fi
1986	352390543	80511462.093	Action, Adventure, Sci-Fi
1988	402111870	80511462.093	Action, Adventure, Sci-Fi
1993	652270625	80511462.093	Action, Adventure, Sci-Fi
2776	165447480	80511462.093	Action, Adventure, Sci-Fi

Genre Predicted Box office revenue, Left, sorted by descending

Gradient Boosted Trees - Predictions

Row No.	Box.Office.Gross ↓	prediction(Box.Office.Gross)	genre
1993	652270625	80511462.093	Action, Adventure, Sci-Fi
1721	534858444	13037318.994	Action, Crime, Drama
267	532177324	80511462.093	Action, Adventure, Sci-Fi
2912	504014165	6161822.339	MISSING
1992	486295561	62498287.249	Animation, Adventure, Comedy
1991	415004880	62498287.249	Animation, Adventure, Comedy
2291	409013994	80511462.093	Action, Adventure, Sci-Fi
1990	404000714	72073646.843	Action, Adventure, Fantasy
1989	402453882	131685325.311	Adventure, Sci-Fi, Thriller
1988	402111870	80511462.093	Action, Adventure, Sci-Fi
2290	389213281	80511462.093	Action, Adventure, Sci-Fi
2631	364001123	32267421.999	Adventure, Drama, Family
3124	363070709	23135077.760	Action, Adventure, Comedy
1987	356461711	62498287.249	Animation, Adventure, Comedy
1986	352390543	80511462.093	Action, Adventure, Sci-Fi
638	337135885	80511462.093	Action, Adventure, Sci-Fi
2289	336530303	38183549.374	Action, Adventure
1720	336045770	51006776.624	Animation, Action, Adventure
1470	333176600	80511462.093	Action, Adventure, Sci-Fi
2288	317101119	72073646.843	Action, Adventure, Fantasy
1719	314057748	80511462.093	Action, Adventure, Sci-Fi
266	303003568	78394869.374	Adventure, Fantasy
1718	301959197	61055269.405	Adventure, Family, Fantasy
637	296623634	123249466.593	Adventure, Drama, Fantasy
536	292298923	123249466.593	Adventure, Drama, Fantasy
1905	291045518	72073646.843	Action, Adventure, Fantasy
265	279261160	17600871.174	Comedy, Romance
1904	268492764	62498287.249	Animation, Adventure, Comedy
1717	262030063	38183549.374	Action, Adventure
535	260044825	24346662.393	Comedy, Family, Fantasy
2630	258766572	80511462.093	Action, Adventure, Sci-Fi
264	258366855	78394869.374	Adventure, Fantasy
634	2582604178	141627361.429	Drama, Fantasy, Crime

Genre actual box office revenue, Left, sorted by Descending

Analysis Findings Summary

Every attempt to normalize results and reduce error measurements made little progress in predicting accurate revenues. As a whole, it appears that the overarching results will lead to little predictable accuracy in factors directly affecting an increase in box office revenue, specifically for lesser performing films. As shown by a root mean squared error of $26729406.044 \pm 2292388.804$ from the gradient boosted trees model. This shows us that the average prediction

was off by \$26.7 million with a deviation of \$2.3 million. It can be stated that each model utilized consistently performed ‘inaccurately’ in predicting an end result for gross box office revenue, however, the overall correlation in predicting which movies will perform well is relatively high with a correlation factor ranging from $0.819 +/ - 0.049$ (82%) to $0.822 +/ - 0.042$ (82%) between the models utilized. This inconsistency in accurate prediction values can be attributed to the rather large values presented to the models ranging up to the high end of hundreds of millions of dollars. So even though the relative error is high, it is an acceptable measure in many ways for box office revenues. It should be said that lower value films with smaller box office revenue predictions will suffer in accuracy.

In this study, a ‘successful’ production refers to the relationship between an independent variable and its relational factor to higher USD values in box office gross revenue.

Within important factors, a ‘PG-13’ rating was the second highest predictor for a successful production, rating only .02 below ‘Budget’ overall which sits at a supporting prediction factor of 0.41 within a gradient boosted tree simulator. In contrast, a ‘Summer’ movie improved predictability by $0.04 +/ - 0.01$, showing that although the season does make minor improvements to predictability, it is not a major concerning variable. With ‘Budget’ listed as the generally most important factor in determining success, it is clear that ‘Budget’ itself is a heavily induced variable due to the fact that a movie’s budget extends to the production studio producing it; thus being able to hire popular actors and directors and providing a substantial marketing budget.

Certain factors such as the ‘metacritic’ and ‘imdb’ ratings can be considered as related to the overall success of a movie, however, it should be noted that these scored ratings themselves are more likely to signify if a movie is a success while not directly affecting or predicting

performance. A higher ‘metacritic’ score is generally associated with more popular movies and is a better signifier of successful movies with a positive weight of 0.249 while an ‘imdb_rating’ only provides a weight of 0.004 showing it has little to no relational predictability.

Topic Questions

(1) Why do some small budget films end up being blockbuster hits? Conversely, why do some large budget films fail?

Some small budget films tend to perform better when they are built upon popular factors like a ‘PG-13’ rating, being released in Summer or the Winter, and being sequels or remakes. In contrast, a high budget film can fail because of doing the opposite. Any ‘Tv’ rating, or those other than ‘PG’, generally performs far worse in comparison to other ratings. Moreover, films released in Fall and Spring will become busts.

(2) Do certain genres lend themselves to higher return? Horror, romantic comedies, science fiction?

Movies that directly performed better included ‘Action, Adventure’ alongside genres like ‘Drama’, ‘Animation’, ‘Comedy’, and ‘Sci-Fi’. Although the predictability of these factors were consistently low in accuracy, at a relative error of 90% between models, it was apparent that movies containing ‘Action’ and ‘Adventure’ performed better in comparison to movies that did not pertain to these two genres.

(3) Do remakes, tent-poles and sequels perform differently?

Sequels tend to perform better, with remakes following closely behind, while tentpoles show no general increase in a successful performance. Although these factors have an impact, they are relatively small overall.

(4) How does the time of year, weather and economic trends influence box office performance?

Summer and winter coincide with an overall prediction in movie success while fall and spring will contradict a successful outcome. The years of release also coincide with times of economic upheavals and health in a general sense. The healthier the working class is, the more money will be available to be spent on recreational activities like attending movie releases. This is an overall major factor considering movies themselves produce income based on a high volume of sales at lower prices.

Interpretations and Inferences

Movies with higher budgets and specified age ratings tend to perform better. Known community favorites will also tend to increase revenue. Following popular trends and providing movies in lower ‘PG’ ratings will draw in larger audiences because it is assumed families with children are more likely to attend new releases in theaters in larger groups than individuals for separate age ratings. Seasonal factors are also more likely to draw in audiences as individuals and families will look for ways to ‘stay out of the weather’. It can be either too hot or cold to stay outside so people will seek to escape the heat or find activities inside to stay warm. Generally speaking, people will find different things to do in nicer weather rather than watch movies in seasons such as the Fall and Spring.

Lastly, a movie with a high budget is likely to have a larger marketing budget and become more widely known than a small budget production film with limited marketing capability. Coinciding with that, a high budget film is likely to be associated with a well known production studio, director, and/or A-list actors, which are all desirable traits for movie

audiences. In this sense, the ‘Budget’ covers a wide range of factors which showcases the value of many traits combined in a singular variable.

Recommendations

It is recommended to prioritize an appropriate movie rating such as ‘PG-13’ or ‘PG’ to properly target major audiences and improve box office performance. Also, movies released in ‘Summer’ and ‘Winter’ are likely to see an increase in performance. In this following order, movies should be prioritized thereafter: sequels, remakes, and tent-pole productions. Sequels and remakes already pertain to fan-bases and communities which draw in larger audiences, while tent-poles, although larger in production, tend to have little impact on overall revenue.

Conclusion

Overall, predicting actual box office gross revenue is a highly convoluted process that includes a number of variables that are hard to account for ranging from cultural acceptance, political environments, current economic success, ticket prices, general health of the working class, and many other unknown factors. So, although major factors were identified in determining whether a movie will succeed or not with a high correlation, accurate predictions for box office revenue are difficult to pinpoint, showcasing a potentially large number of unpredictable and indeterminate factors affecting gross revenue. For now it is important to follow the discovered insights to increase the overall chances of improving success within the film industry.