4/24/2018
Brook Bi

NBC Universal - Predicting Success in Theaters

Introduction

What makes a movie "one of the greats"? While that question may elude both data analysts and artists for years to come, this analysis points to some effective indicators of box office success for films, in the hopes of maintaining a thriving movie industry for years to come. This project focuses on several commonly cited reasons for movie fiscal success, and analyzes how much of an impact each factor has on the bottom line.

This is a significant issue, because even while projections have 2017 as a good movie year, industry revenue is down 10% from 2016. This analysis will propose solutions for what a studio can do to change the direction of these trends, as well as a predictive model for whether a new movie that comes out will be profitable or not. NBC Universal or other movie studios can apply these findings both on a large scale (in choosing genres and directions to focus on) and on a small scale (in directing and producing individual movies that are profitable.

The datasets used for this analysis come from NBC Universal directly, as well as several assembled lists from online polls (e.g. "Greatest Actors/Actresses of All Time" - Ranker.com and "Greatest Directors of All Time"), having over 1 million and half a million votes, respectively.

Profitability was the defining metric used for analysis in this study - using Python to clean the empty and irrelevant cells, a "profitability ratio" was determined for each movie based on its box office sales divided by budget (and controlling for foreign films and currencies). In assessing the value of name-brand recognition in Hollywood, I crossed a list of the "greatest" actors and actresses of all time, assigning a count to each movie to later plot against profitability. For both this and the ratings question (data provided in the challenge) I created a regression model to "learn" to predict movie success as a factor of famous actor/director presence and online ratings.

While there are certainly limitations to this data set and analysis (e.g. limited number of data points for machine learning, inaccurate labeling and biased focus on foreign films, and subjective descriptions on the movies, this should certainly serve as a tool and starting point for aspiring and current filmmakers.

The products of this analysis include the code and final datasets, a written summary of the big results, as well as an infographic summarizing the results to the average layman.

4/24/2018
Brook Bi

Treating of variables for Machine learning

Treating and Cleaning

As is standard with raw data sets, cleaning was an essential pre-processing step to ensure valid and applicable results. Information in columns such as a movie's ID on website IMDb, a short description of plot for each movie, and date of DVD release were excluded from analysis. Rows containing empty values, dollar amounts in foreign currencies, and text in numerical fields (e.g. "gross but totally worth it" as a Budget value) were treated and ultimately excluded from analysis.

In addition to cleaning, additional variables were derived to better answer the question at hand. The most important derived variable, profitability ratio, was defined as gross budget divided by gross sales for a particular movie - thus, any movie without a value in either column was discarded for the sake of regression. This ultimately became the target variable in the regression analysis. Another variable was created and considered for this role; gross profitability (box office sales minus budget), however, did not effectively control for the magnitude difference between big budget studio and small, independently released movies. Figures obtained for this study were not adjusted for taxes or inflation, which also better justifies the use of profitability ratio instead of profitability itself for movies coming from different time periods.

Categorical variables (e.g. genre and MPAA censorship ratings) were converted into dummy binary variables to enable quantitative analysis. For genres, movies that fell were categorized as multiple genres were treated inclusively, giving an affirmative value to multiple columns of genres, rather than trying to determine the "best" genre to place a movie in. For MPAA rating, movies only had one MPAA rating.

Budget was initially used as-is in the regression model, but the presence of a few but extremely profitable/unprofitable outliers hindered the efficacy of the model as a whole. As a result, the "floor", or lowest written value, for budget was set to $1.1M and the "ceiling", or highest written value, was set to $160M in order to make for a more standardized and predictive regression model. The values were determined by analyzing the distribution of budget sizes and retaining the median 90% of budgets.

Actor and Director Score

In order to answer the question of whether hiring famous actors and actresses really attracts significantly more customers to a movie, two variables - "actor score" and "director score" - were derived and used in regression analysis.

Actor score was a created variable made by cross-checking actors listed in the dataset across a Ranker Community poll (containing 1.1 million crowdsourced votes) and assigning a "point" for each top famous actor listed in a movie's info. To help account for famous but perhaps controversial or less well-recognized actors, the list of 100 was supplemented with IMDB's own Top 100 Greatest Actors of All Time (curated by website staff) list. Scores ranged from 0 to 3 points, as no movie had four or more qualifying actors.

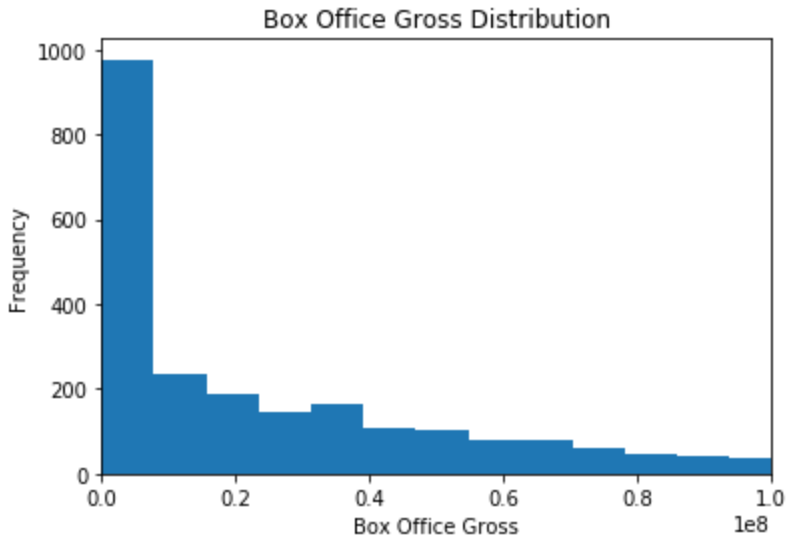
Director score was determined similarly, using Ranker's director crowdranked list (142 thousand votes) to assign scores to movies for famous directors that cross-checked with the movie's data. As there is typically only one director for a movie, the scores were tiered according to just how famous a director was on the list (directors ranked 1-10 received a max score of 5 points, 11-20 received 4 points, etc.).

Pre- and Post-release Models

A post-release linear regression model containing all treated variables was created for this analysis. From the post-release model, any variables that might not be available during a movie's initial release in theaters, including IMDB rating, were removed to leave a "pre-release" regression model. Note, however, that a growing number of websites today (e.g. IMDB and Rotten Tomatoes) are releasing preliminary critic and audience interest scores before a movie is released.

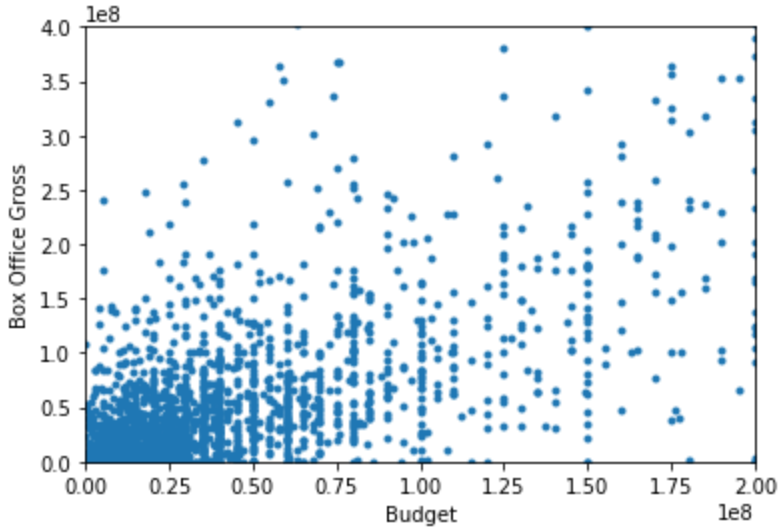
Descriptive & Inferential Statistics

The distribution of both box office grosses follows an exponential decay pattern, with most movies producing very little money.

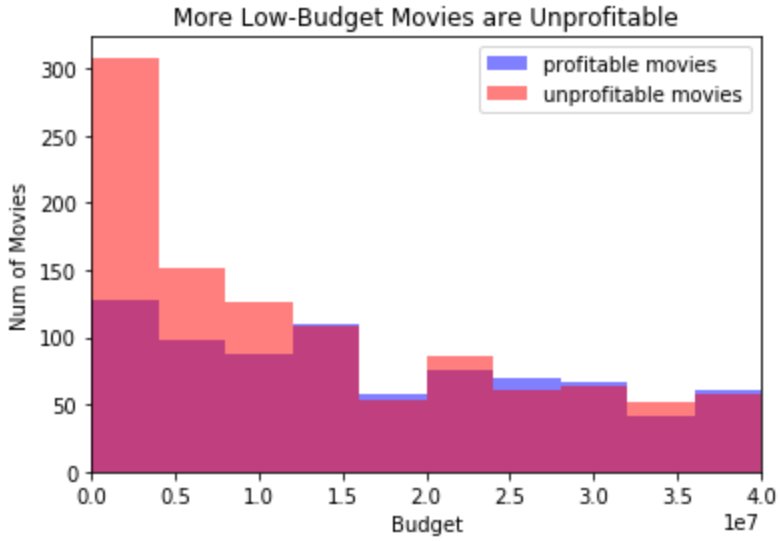


The vast majority of movies return very little in the box office.

The median movie has profitability of 0.809, meaning it has a return of 80.9 cents on the dollar. However, the mean profitability is over 1.2, suggesting that some movies are vastly more profitable than the rest while the majority are slightly unprofitable. The scatter plot reveals that, in general, a higher budget results in a higher return. However, there are many exceptions to the rule.



There is a defined correlation between size of budget and size of box office gross - however, the distribution of points suggests that there are many movies that stray far from the best fit line.

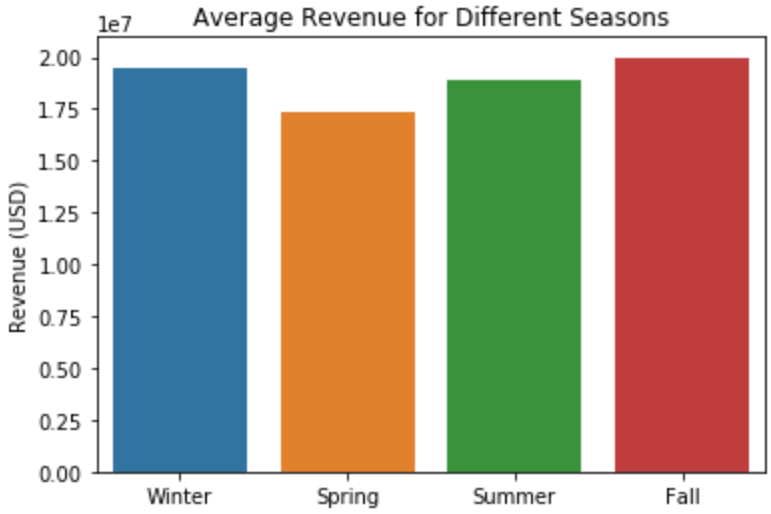


Magenta coloring represents the overlapping space between profitable and unprofitable movies. Unprofitable movies follow an exponential decay pattern, with many low-budget movies not succeeding. Profitable movies have a much flatter distribution.

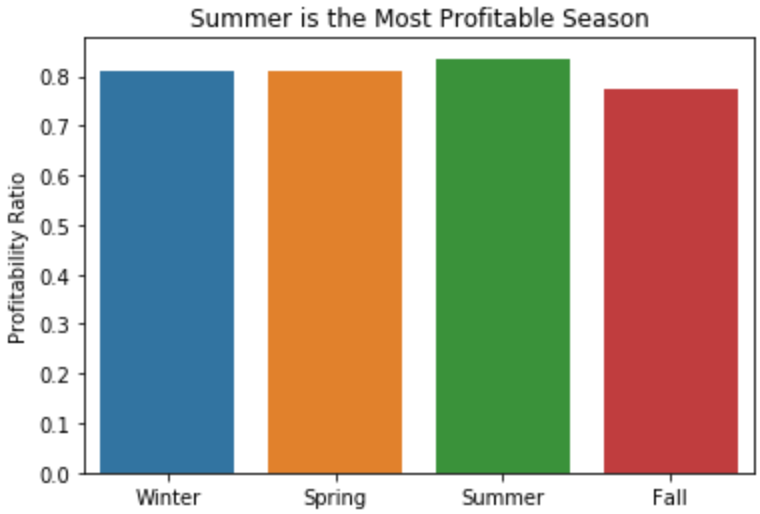
Comparing the distribution of profitable to unprofitable movies, it becomes clear that the unprofitable ones are responsible for the decay shape of the graph. When budget size is $15M or greater, there is a roughly even number of profitable and unprofitable movies. This suggests that budget size could significantly influence profitability.

Seasonal analysis

Based on release date, movies were grouped according to season to determine if certain times of the year are better to release than others. While movies in the fall have the highest average gross revenue, they also cost the most. The most profitable season is actually summer.

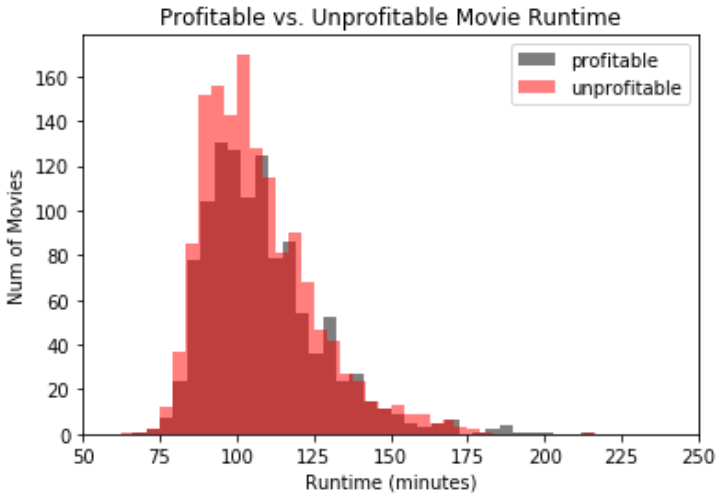


While Fall movies have the highest average revenue...



Movies in the summer actually have the highest return on investment, on average.

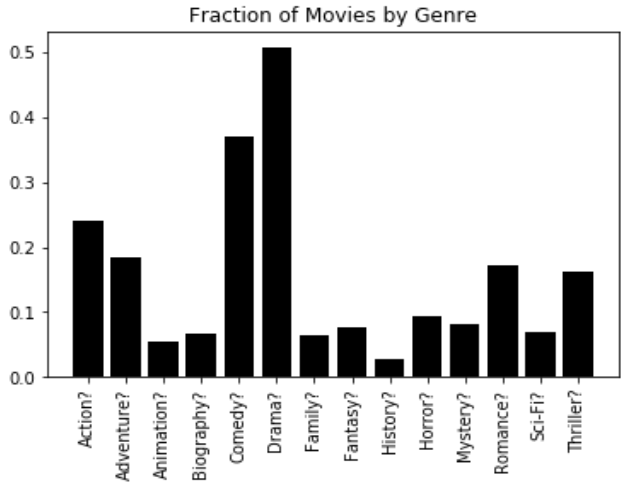
Runtime analysis



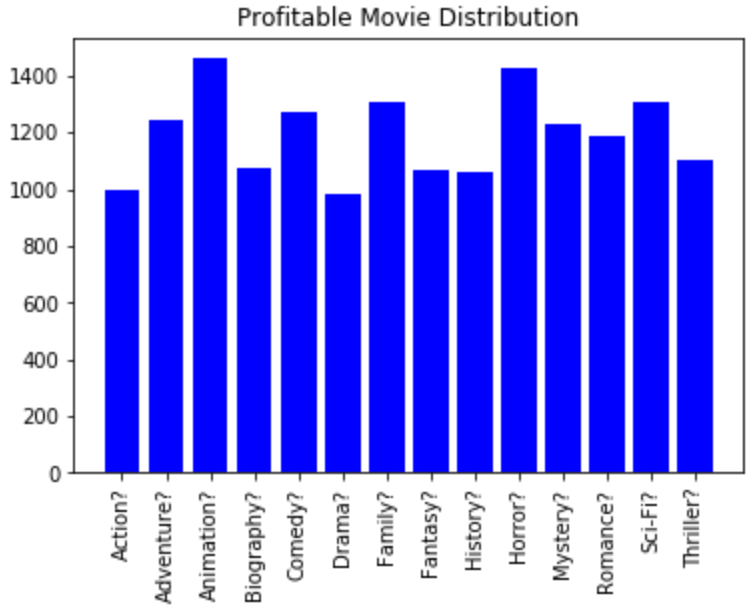
Runtime is distributed in a distribution leaning more heavily on the left. While the median movie runtime is 104 minutes, several exceptionally long movies make the mean 107. At first glance, it looks like longer movies slightly outperform shorter ones, based on the distributions of the two graphs. The darker pink space (which represents shared space) is noticeably shorter than the unprofitable distribution, suggesting that runtimes in the 80-110 minute range may be riskier. However, this could also be noise resulting from a relatively small sample size of movies.

Genre Analysis

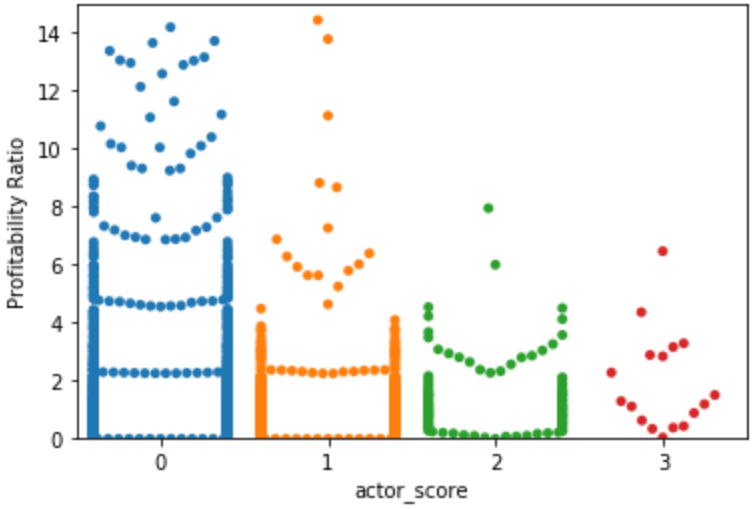
Comedy and drama are easily the most common genres released in theaters, with frequencies both over 0.3 (3 in 10 movies qualifying). However, in terms of profitability, clear winners are not as easily visible. Animation and Horror actually have the most profitable movies that qualify, but not by much.



Comedy and drama are shown to be the most common genres, but the actual profitability of different genres is more complicated than that. Further analysis was performed on each genre by assigning each genre a dummy variable and testing for significance.



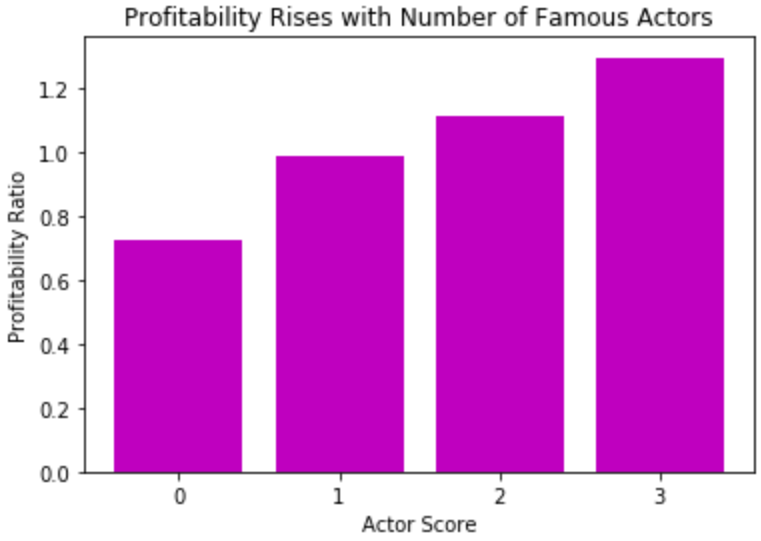
Actor and Director Score



As might be expected, most movies don't have the privilege of featuring a famous actor, let alone two or three.

As discussed earlier, an actor score (number of famous movies) was determined for all movies in the dataset. The swarm plot above shows that having more famous actors tightens the range of profitability, most likely due to the increased size of budgets needed to hire more famous actors. At first glance it appears that movies with more famous actors are more profitable - however, the small data size

of movies with two or three all-star actors may make this hypothesis difficult to accurately assess without regression analysis.



Statistically, mean profitability increases with actor score (number of top-100 actors). Movies with two or more "actor points" have a median profitability above 1.

Machine Learning Statistics

Multiple Linear Regression was the primary method used to evaluate the effect size of each variable hypothesized to influence profitability. The target variable for representing profitability was profitability, because it best manifests the question of how one can predict whether a movie will be financially successful.

The data was divided into a training set, validation set, and test set in a 60/20/20 split. There were two main models created - a pre-release, predictive set (excluding several variables) and a post-release, adaptive set. In today's environment, however, the post-release model can likely be used for prediction, given the increasing availability of movie ratings before their release date. The MSE change for dropping certain variables was analyzed, and many variables that appeared to be significant during the descriptive and inferential testing actually proved redundant.

Regression was used to create models for prediction, and OLS analysis was used to calculate the p value for each variable. Variables with an insignificant p value ($p > .05$) were filtered out of the next iteration of each model. This process was repeated until many of the original and dummy variables were eliminated, and all that remained were variables that bore a significant p value.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.8871	0.697	4.141	0.000	1.519	4.255
Budget_adjusted	-4.331e-08	6.28e-09	-6.900	0.000	-5.56e-08	-3.1e-08
runtime	0.0073	0.007	1.060	0.289	-0.006	0.021
Drama	-0.9363	0.260	-3.596	0.000	-1.447	-0.425
Horror	1.8740	0.413	4.542	0.000	1.065	2.683

Analysis of an OLS model. 'runtime' has a p value (0.289) much higher than our threshold (p < .05), so it is discarded and the model is run again.

After selecting for significant predictor variables, the validation set was used to help protect against overfitting. Models with different combinations of variables were compared against one another using the validation set and MSE (mean squared error) as a metric for quality. The train-test-split method was used even for a relatively simple method like linear regression . This is because a large number of dummy variables compared to the number of samples increased our risk of overfitting with our models.

The most significant pre-release variable was budget, which was treated by flooring and capping its most extreme values. Overall, this - though only having an ($R^2 = 0.07$) or so - was the best predictor in the pre-release model. The most significant post-release variable was IMDB rating, a 0 to 10 metric by which online critics (generally crowdsourced) vote on the overall quality of a movie. However, an IMDb rating is a **weighted** average of votes cast online, meant to protect against vote stuffing but weighted according to an undisclosed method.

Results

Based on initial OLS modeling on the training set, the Budget (adjusted), Horror, and Drama variables appeared to be the significant variables in the equation. When analyzing the validation set with MSE testing, only the variables Budget and Drama improved the quality of the regression model. With budget already accounted for, drama movies **underperformed** other genres in terms of profitability, while it remained unclear whether horror movies truly had a significant effect on profitability.

Drama was significant in the training dataset, but its effect size was small, only slightly worsening the MSE when removed from the validation model.

Horror was a somewhat different variable, seemingly also proven statistically significant in the training data set (through OLS analysis). Yet while it was significant on its own, MSE (mean squared error) testing revealed that it did not contribute a large amount to the accuracy of the model in conjunction with other variables. Removing it actually created a more accurate model in both the pre-release and post-release versions of the validation data set.

Given the small dataset of movies, Horror's predictive ability (or lack thereof) may have been a false positive in the training set or a false negative in the validation set. While further testing needs to be performed to determine with greater confidence, the matter of Horror's significance could affect the

recommendations for movie producers. These two cases will be discussed in the *Recommendations and Conclusion* section.

There were many variables that did not prove to be significant. These include the remainder of the genre variables, any of the MPAA ratings, runtime, how long ago the movie was released ("age"), season of release, and fame of actors or directors present within the movie.

It was surprising that even though the inclusion of famous actors and directors did correlate with increased profitability, the regression analysis indicated that both variables did not have a significant impact on the target variable. This likely indicates that the effect was mediated by another variable, possibly because larger budget sizes can afford more actors. Also interesting was the fact that fall, despite being the most profitable season, did not come through as a significant season in the research - this could have been due to other underlying variables that were not explored, such as also being influenced by budget size.

Recommendations and Conclusion

Based on the research presented in this workbook, there are several clearly defined steps and processes that NBCUniversal can take to maximize profits amidst a challenging industry, as well as several future directions for research moving forward. We can't say that it's better to invest in big-name actors and directors. We also can't recommend releasing a movie geared towards children ('G' rating) vs. an adult or un-rated movie ('R' or 'NR'). We can't recommend most genres as being better than others, or longer vs. shorter movies. However, there are certain business decisions that do seem pertinent and promising, based on the data.

In terms of the results on budgets, the solution would be relatively simple for big studios like NBCUniversal – concentrate on releasing quality big-budget movies, and ensure directors get proper funding for the movies they produce. For smaller movie-releasing studios and people, our advice is to pursue crowdfunding or sponsorship if at all possible, because a well-polished and positioned movie is worth the budget according to the data.

Another simple change the studio could make would be to release a smaller proportion of drama movies. Furthermore, future research can analyze features of drama movies to better understand what may make them less successful. Could it be related to the sheer number of drama films being release? Clustering and unsupervised learning can be used to break apart the distinguishing features of drama movies and discover what isn't working.

As for horror, there are two cases and corresponding recommendations for the results obtained by the study. If the validation results were incorrect (perhaps due to the small data size) and the variable is significant, then the same kind of analysis should be performed on horror as with drama movies. With two genres to cluster, more robust and confident insights can be determined regarding factors that make movies more successful. However, if additional research reveals that the test set results were incorrect, then the general recommendation would be to not emphasize horror movie production and treat the result as overfitting.

As evidenced by our post-release model, quality is paramount in predicting which movies will ultimately help a studio's bottom line. Follow-up studies can be done to assess which rating system may have the most value (i.e. comparing IMDb to other movie rating websites, like Rotten Tomatoes and Google Play), but overall the evidence suggests that, in the long run, critics and reviews can pick out winners from losers. A future model would improve in value if it could predict how humans might rate it, and one idea for this would be to utilize NLP analysis of movie scripts to determine if certain keywords, features, or metrics can predict movie ratings.

Other future studies could also delve deeper into the famous actors questions, plotting famous actors vs. budget size to see if these correlate to prove that budget size mediates for the effect of famous actors. They could also gather more movies with 2 or 3 famous actors, since we had a prohibitively small number of data points in these categories. Furthermore, dividing director score into 5 tiers of points may not have been the most accurate approach, and future studies could also approach this variable differently. One last idea would be to study whether movies that have a greater amount of time between theater and DVD releases are more successful (due to perhaps having more time to "simmer" and make money in the theaters).

Ultimately, the best move forward would be to delve deeper into the questions and extensions that sprung up in this analysis to gain an even better understanding of what the story is behind the data. One of the greatest weaknesses in these findings is the size (or lack thereof) of the dataset. Either by researching and adding on the budget and gross figures or by obtaining a separate dataset of entries, we can attain greater confidence in our results with a larger sample size. Before results from these subsequent studies come in, there are multiple measures NBCUniversal can take now to improve outcomes for the 2018 movie year.