PREDICTING BOX-OFFICE SUCCESS OF MOVIES IN THE U.S. MARKET

Darin Im and Minh Thao Nguyen

CS 229, Fall 2011

I.   INTRODUCTION

The movie industry is a multi-billion dollar industry, generating approximately $10 billion of revenue annually.[1] In recent years, movies have generally become divided into two categories: blockbusters and independent movies. Studios have focused on relying on only a handful of extremely expensive movies every year to make sure they remain profitable. It is estimated that 80% of the industry's profits over the last decade is generated from just 6% of the films released; 78% of movies have lost money of the same time period.[2] These blockbuster movies emphasize the spectacular: casting as much star power as possible and pairing it with high production value. The result of this is a sky-rocketing budget. It is estimated that the average movie now costs $100.3 million after including production and marketing expenses.[2] However, "Hollywood is the land of hunch and wild guess," so it's difficult to predict whether these high-budget films will actually make a profit.

As the costs of movies have gone up, it has become paramount that movies are successful to justify such large undertakings. Studios are under great pressure to ensure their movies succeed, trying to find ways to produce movies that are more likely to be successful. However, this is much hard said than done. Jack Valenti, President and CEO of the Motion Picture Association of America (MPAA) once said, "No one can tell you how a movie is going to do in the marketplace. Not until the film opens in darkened theatre and sparks fly up between the screen and the audience."[2]

Because of this the movie industry has attempted to employ the help of computer scientist to create recommendation and predictive software to tackle this problem. Recommendation software is more common and attempts to make correlations between a consumer's past choices and other products they might like. The recent Netflix Prize competition caused a great surge in the creation recommendation algorithms by providing $1 million prize to anyone that can improve their algorithm by 10%. Predictive software is less common, and typically highly inaccurate. Prediction software attempts to predict the success of a movie using only the details known pre-release. This project attempts to employ machine learning to predict the expected profits of a movie.

II.   DATA

The data used for this project was obtained from IMDb using a Python script to scrape the data. We limited the movies we used in our project to feature films released between 2001 and 2010, were in English, and had a gross revenue of at least $500,000. Using these search criteria we were able to find 1,937 movies.

We had a number of reasons for placing these restrictions. The reason for placing the time period restriction was that we only wanted to include recent movies as it would have been difficult to compare movies from different eras. Over time, movie tastes would have changed, meaning the characteristics of a profitable movie would have also changed. Our model would be unable to take these changes into account. This time period restriction was also placed so that gross revenue comparisons wouldn't be significantly impacted by the rates of inflation. If this needed to be taken into account, it would have further complicated our model. Also, we only included movies that were in English as we were only interested in U.S. box office results and thus were only looking at U.S. gross revenue. Foreign films were often reported in Euros, necessitating that we take the conversion rate into account. Lastly, we required that all movies have a gross revenue of at least $500,000 to reduce the number of independent movies that have a limited amount of information available.

III.   FEATURES

The features that we included were the genres (as classified by IMDb), the user rating, the number of user ratings, the budget, the run time, the MPAA rating (G, PG, PG-13, etc.), the studio company that produced the movie, the number of screens during the opening weekend, and the time of year the movie was released.

All of these features are available pre-release, with the exception of the average use rating and the number of user ratings. The reason we decided to include these features into our model was to determine whether they would significantly affect the projected revenues.

In King's paper, it was found that there was no apparent correlation between box office success and critic approval.[5] This is surprising given the fact that there are a number of movies available for viewing and movie-goers only see a select few, requiring them to rely on some source to determine which of these movies are worth watching. Because of King's findings, we were curious to see whether user ratings have a significant impact on revenue. According to Kennedy's paper, there is a fairly strong relation between user ratings and critic ratings, meaning movie goers generally agree with a critic's opinion. However, movie's that had critical acclaim didn't necessarily do any better, raising the question of whether movie goer's still see movies they anticipate them to be "bad."

IV.   MODEL

To obtain a projected revenue for a movie, we implemented a linear gradient descent algorithm. Features such as genres, seasons, MPAA movie ratings, and studio production companies are binary classifications that can be assigned a value of 0 or 1. However, our other features such as the budget, number of ratings, and user ratings are not binary classifications and have large values that can be highly variable. Because of this we normalized the features so that their values have a mean of 0 and variance of 1. We accomplish this by computing the following 4 steps:

1. Set $\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$
2. Replace each $x^{(i)}$ with $x^{(i)} - \mu$

3. Let $\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x_j^{(i)}\right)^2$
4. Replace each $x_j^{(i)}$ with $x_j^{(i)}/\sigma_j$

Having calculated a predicted revenue, we will classify them into profit buckets to get an idea of what range these values come in. For our model we have classified 9 distinct buckets:

1	2	3	4	5	6	7	8	9
< $1 M	$1-$10M	$10-$20M	$20-$40M	$40-$65M	$65-$100M	$100-$150M	$150-$200M	> $200M

This is a similar approach that Sharda took in his model,[3] which seems appropriate for the task at hand. Although it would be nice to have a model to predict the exact amount a movie will make, studios would likely be satisfied with knowing what sort of category the movie would fall in. Will the movie be a "flop," hardly making any money and costing the studio company? Will the movie at least be able to at least make enough to cover its costs? Or will it be a "blockbuster" and bring the studio an embarrassment of riches?

We can also approach our bucketing method by a modified version of k-means clustering. When we calculate the centroid of each cluster (the buckets), the centroid will be heavily determined by the gross revenue since all of the other values other than the gross revenue has been normalized. Then to assign each data point to a cluster, we will find the mean of each cluster and find the cluster that is closest to each data point/

In addition to predicting the gross revenue of a movie, we would also like to determine what the most important attributes of a profitable movie are. This is especially important since we included two features—user rating and number of user ratings—that are not available pre-release. Although, PCA provides a method to map our data set to a smaller subspace, it may not select any of the components to be the basis of this new subspace. To accomplish our task, we devised a "Variance Minimization of the Gross" process, a variation of the PCA method. In our VMG process, each component is removed from the data, and then the data is tested without that component. The most important components then are the components that have the largest sum of the squares of the differences between the new predicted gross and the old predicted gross.

V. RESULTS

Similar to the study conducted by Sharda, we used k-fold cross-validation rather than running a single experiment, specifically 4-fold cross-validation. Since the number of movies we had was not divisible by 4, we simply removed 1 movie to make our calculations slightly easier.

The result we found was that on average our estimate was off by approximately 57.1%, a rather high error. However, our model was more successful in predicting whether a movie will be profitable of not (i.e. whether the projected revenue will exceed the budget for the movie). When determining the profitability was the objective of our model, it was able to correctly predict this approximately 72.4% of the time. Therefore, although our model provides a rather large range of

what the projected revenue may be, it is at least useful in determining whether a movie is worth producing (i.e. whether a movie is expected to be profitable).

When trying to assign movies to revenue buckets, when found that we were able to assign a movie to the correct bucket approximately 25.3% of the time. Additionally it was found that we were able to assign a movie within 1 class of the correct bucket approximately 52.1% of the time. When using our modified k-means clustering method, we were able to correctly assign a movie to the correct bucket 29.6% of the time and be within 1 bucket 57.6% of the time.

Additionally, for our VMG (variance minimization of the gross) process, we found that the two most important features were the number of screens during the opening weekend as well as the season it was released. This would imply that our assumption that user ratings and the number of user ratings shouldn't significantly affect the predicted revenue was fairly safe. However, we did find that they do vary our results to an extent (they were among the top half of the components we used). Therefore it would be safer to exclude these features from future models.

VI.   COMPARISON TO OTHER MODELS

Although our model's ability to predict revenue is not highly accurate, the results are in the same ballpark as other similar studies. In particular, Sharda cited results of being able to correctly classify a movie in its appropriate revenue bucket only 36.9% of the time and that it was within one bucket 75.2% of the time. Sharda claims that his results exceed those of other studies that have been conducted.[3] Additionally, Chen found highly variable results as well. In Chen's model, he provides a wide range of revenue. For example, for the release of Hannibal he predicted a low of $19 million and a high of $1 billion, a range that spans almost $1 billion.[8] Furthermore, the predicted revenues that he found were also off from the actual revenue. He claims his prediction for Joe Dirt to be fairly close, but his estimated revenue was $29.9 million while the actual was $22.7 million, a 31.7% error.[8] Because of this, we believe that our results do not deviate much from the typical results of this sort of model.

Though it is expected that our results are low, they are still lower than Sharda's results and there are several attributes of our model that would suggest that our model can be improved. For one, the model's ability to accurately predict movie profitability may be affected by the fact that it assumes that the successes of individual movies are independent of each other. This is a very strong assumption. For example, our model won't take into account movie sequels that would exceed typical movie expectations because of the strength of the franchise or that movies will have lower revenues if the quality of the other movies opening that weekend is better.

Additionally, we did not take into account the "star power" of movies like Sharda did. Sharda graded each movie, assigning them a letter from an A+ to a C based on the based on the successes of the actors/actresses in the movies.[3] \ However, we believe that it would be difficult to implement Sharda's method since it requires intimate knowledge of each movie (he assigned each movie a grade). This may be the most lacking feature in our model since Simoff and Sparrow found that their "star power" variable had the strongest effect on revenue out of all their features.

There were also features that Sharda included that we were unable to include because it would require intimate knowledge of each movie. In addition to grading "star power," Sharda also graded the technical effects as well as the competition from other movies.[3] These likely increased the accuracies of Sharda's results, but we were not able to include these in out model as it would require extensive research to assess the level of competition and grade the technical effects of each movie since we were unable to find sources that had such information readily available.

VII.   CONCLUSION

The film industry is a risky business. As screenwriter William Goldman once said, "nobody knows anything."[5] Although it is becoming increasingly important for studio executives to be able to accurately predict movie revenues before they are released, the consensus by most studies is that most attempts are still inexact; the prediction is of movie revenues are still more accurately classified as an art rather than a science with most experts predicting revenue based on rules of thumb, hunches, and their experience.[3] Although our model isn't likely to be snatched up by movie executives anytime soon, it appears that predictive modeling of movie revenue in general is still a work in progress.

However, we do believe that it would be possible to increase the accuracy of our model from its current state. Perhaps the most important change that we can make is to implement some feature to take into account the "star power" of a movie. Additionally we could take into account other features such as the origin country of the movie, writers, directors, and whether it was released on a holiday weekend.

VIII.   REFERENCES

[1]   "The Numbers - Movie Market Summary 1995 to 2011." *The Numbers - Movie Box Office Data, Film Stars, Idle Speculation*. Web. <http://www.the-numbers.com/market/>.

[2]   Davenport, Thomas H., and Jeanne G. Harris. "What People Want (and How to Predict It)."*MIT Sloan Management Review* 50.2 (2009). Web.

[3]   Sharda, Ramesh. "Predicting Box-Office Success of Motion Pictures With Neural Networks"

[4]   The International Movie Database (IMDb). <http://imdb.com/>

[5]   King, Timothy. "Does film criticism affect box office earnings? Evidence from movies released in the U.S. in 2003."

[6]   Kennedy, Alec. "Predicting Box Office Success: Do Critical Review Really Matter?."

[7]   J. S. Simonoff and I. R. Sparrow, "Predicting movie grosses: Winners and losers, blockbusters and sleepers," Chance, vol. 13(3), pp. 15–24, 2000.

[8]   A. Chen, "Forecasting gross revenues at the movie box office," University of Washington, Seattle, WA, June 2002.