Group 2

Andreana Able
Sakol Chompoonich
Chu-Tzu Ko
Kumiko Kotera
Thomas Lotze

Understanding the Factors Creating a Box Office Hit

Examining Movie Traits to Breakeven

Executive Summary

Producing films is one of the riskiest of businesses. A 'mega-project' movie (e.g. Superman Returns) can cost up to $270 million and take years to produce. To justify this investment, the mega-project movie has to make its way to the top 5 of the box office for the year, and generate approximately $300 - $400 million total gross revenue. In order to give some insight into which movies justify their investment in time and capital, we look at whether a movie breakeven simply from box office sales. If a movie makes more total gross than was spent on budget, it has broken even. This is a prime indicator for industry analysts in determining whether a movie is valuable to the studio, since in addition to box office sales, there are significant profits to be made from DVD sales and licensing; if box office sales breakeven, the total profit is significant.

To determine whether a movie will breakeven, we have analyzed the past three years of movie data for major motion pictures. We find that the following factors are significant:

- **Distribution is key**. In order to breakeven, a movie must make it into a large number of theaters.
- **Opening weekend success is important**. Opening weekend performance is already widely used in the industry to predict success. We merely confirm its validity here.
- **High budget means high risk**. Again, supporting industry wisdom, we find that high budgets, while they potentially offer higher reward, are less likely to breakeven from box office sales.
- **Quality still counts**. Quality (as measured by IMDB rating) improves the chance to breakeven.
- **Seek the season**. Being a Summer or Christmas release (when there are more moviegoers and fewer competing movies) helps.

However, there is still a large amount of unexplained variance in the breakeven explanation (an overall error of 11%). While there are some factors which can help explain which movies become hits and justify their return on investment, our analysis shows that there is a significant role that must be played by producers and movie executives. Although these predictors can help focus their analysis, a large part of what makes a movie successful is not captured in these simple quantitative variables; being able to distinguish the elusive quality of a "Pirates of the Caribbean" from a "Gigli" is what makes a good movie executive so valuable.

Technical Summary

Data Description

Our initial data set comes from www.the-movie-times.com, which lists the movies which spend at least one week in the U.S. box office top 60. This covers nearly every major

motion picture which airs in the U.S., with the exception of extremely small-run independent movies. For every notable movie production company, all of its movies will spend at least one week in the top 60. Although some of the extremely small movies may make back their production costs, they are not relevant to the business decisions made by a major movie production company, and so we do not consider them.

The original data consists of Movie name, Release Date, Opening weekend U.S. Gross Revenue (Opening Gross Revenue), Total U.S. Gross Revenue (Total Gross Revenue), Number of Theaters, and Number of Weeks in Box Office Top 60. For this analysis, we use all movies released in 2004, 2005 and 2006, which consists of 1,008 movies, as representative of recent movie performance. Moreover, we retrieved additional movie data from www.imdb.com, including Primary Movie Genre, MPAA Rating, IMDB Movie Rating, Number of Voters, Running Time, Estimated Production Budget, and binary variables indicating whether the movie is a sequel or has a sequel.

Data Processing

Using this, we created several derived variables for each movie:

- **Breakeven**: the target variable, a binary variable representing whether a movie made back its production budget in total box office gross. This is an important industry measure separating strong performers from weak ones. Most recent movies (66.96%) do not breakeven from box office sales alone, and require additional revenue (such as DVD sales) to recoup production costs.

- **Binned movie genres**: Initially, the data consisted of 26 categories of movie genre; we binned similar types of movie genre into 6 groups, based on our domain knowledge and data characteristics. For example, we combine Action, Adventure, and War to “Action.

- **Summer release** and **Christmas release**: We created 2 dummy variables based on movie release date; one for Summer Release (movies that are released from May through August) and Christmas release (movies that are released in December). Based on our domain knowledge, we believe those two are the peak period where audiences go to the theater and big budget movies are released.

- **Restricted movie rating**: We created a dummy variable for any movie with MPAA rating “R” and “N-17”, as these movie ratings seriously restrict the audiences and might impact the profitability of the movie.

Data Exploration

We used several methods to explore the data; example graphs of these methods can be seen in Appendix A. One main method was to use boxplots to determine variables which showed strong separation on the breakeven variable. A second method was to consider scatterplots showing relationships between two of our predictor variables. We also used moving average and other smoothing methods to examine seasonal patterns in movie gross and number of movie releases. Using these, we were able to get a sense for which variables were related to each other, as well as validate those which we expected (from industry analysis) to be important for movie success. For example, OpeningGross and Total Gross are strongly positively correlated, which means that OpeningGross can be used as a strong predictor of Total Gross.

Model Results and Conclusions

Based on our data exploration and information from industry reports, we performed a logistic regression on Breakeven, using the variables that offer significant classification power: “IMDB Rating”, “Budget”, “Opening Gross Revenue”, and “Number of Theaters”. The model offers acceptable error rate at 12.10%, multiple R-squared is 0.5249, and every variable is significant (p-value less than 0.05). We also attempted Discriminant Analysis and Classification Tree analysis, and while they were instructive in helping us validate good predictors, we feel that the Logistic Regression model provides the most accurate insight into the factors involved in a movie’s success.

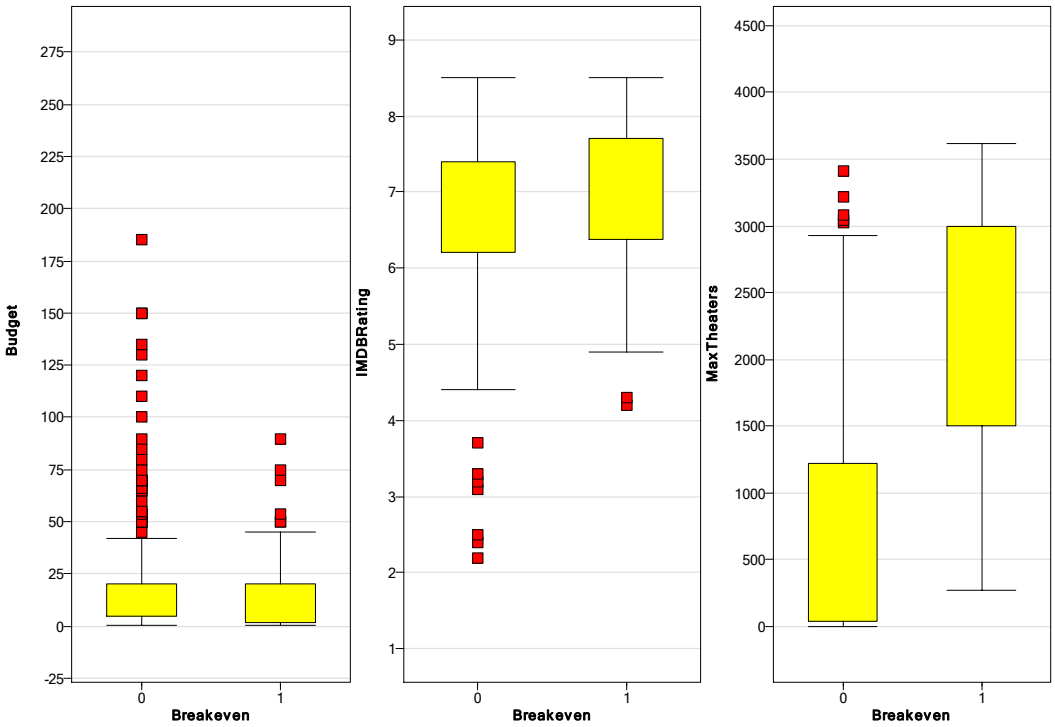
We decided to continue with Logistic Regression Method and include more variables based on our domain knowledge. The second logistic regression model includes the additional 2 dummy variables “Summer Release” and “Christmas Release”. This model has lower overall error rate at 11.11% and supports our exploration of the movie factors. We also explored the use of MPAA ratings and sequel indicators, but surprisingly found that these were not significant in the presence of other variables. This seems to indicate that while they are related to other factors which correlate with movie success (such as opening box office), they are not directly related to movie success, only through their correlated factors. Also surprising was the fact that genre is not correlated with success, given the imdb rating. Although some genres do tend to breakeven more often (specifically Action/Adventure, Animation, and Crime are all significant on their own), it appears that those genres simply tend to have higher quality, and so their improved performance is due to this correlation and not an actual genre effect. There does not even seem to be a seasonal interaction with genre (based on our modeling whether certain kinds of movies do better at different times of year).

Consequently, we use the Logistic Regression Model that includes IMDB rating, Budget, Opening Gross Revenue, Number of Theaters, Summer Release, and Christmas Release as our best model. This model indicates several things. First, that a high quality (as measured by IMDB rating) is valuable. Also, a high budget increases the risk of not breaking even. Third, the model seems to confirm that opening gross is a good predictor of overall revenue (and hence of breaking even). Fourth, the number of theaters is extremely significant in whether or not the movie is breakeven. Finally, Summer and Christmas seasons are related to breaking even.

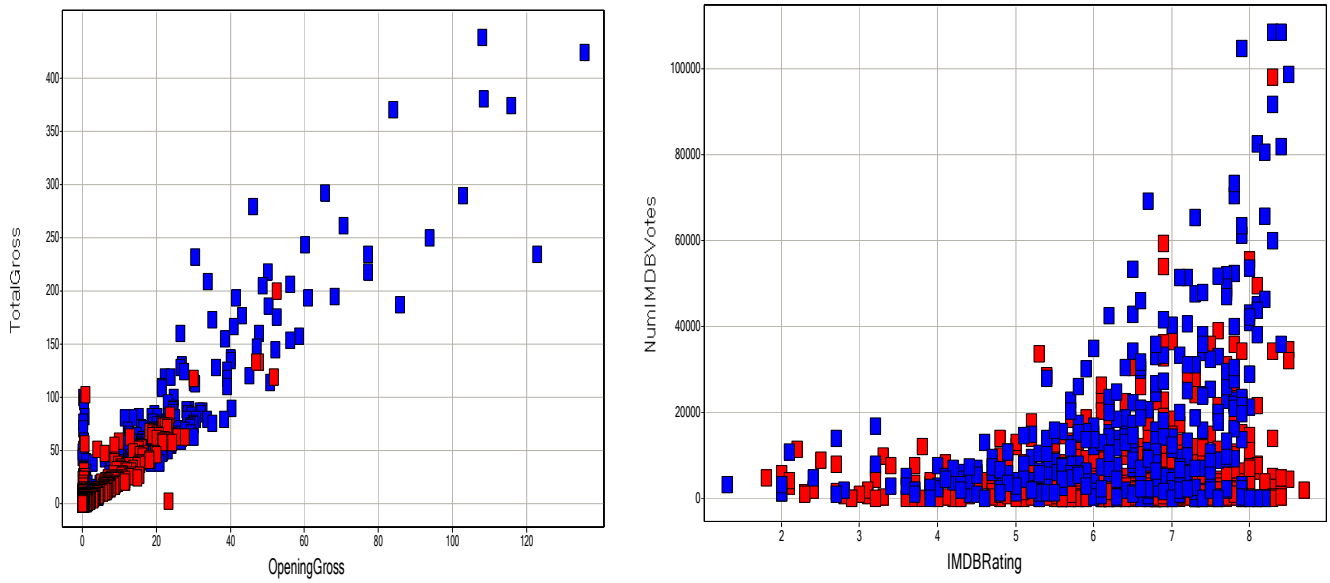
In their paper “Predicting movie grosses: Winners and losers, blockbusters and sleepers”, Simonoff and Sparrow also see some of these same effects, such as the use of opening gross to predict total gross and the predictive power of some genres. It is interesting to note, though, that while they find that budget is a positive predictor of total gross, we can see from our analysis that it also brings with it an increased chance of not breaking even.

The factors identified here show interesting explanatory power for whether a movie will breakeven. However, we also see a relatively high error rate (11.11% overall), especially on movies which do breakeven (25.83%). From this, we conclude that there are still important explanatory movie qualities which are not quantified in our model.

Appendix A: Exploratory Graphs



These boxplots show separation of breakeven on budget, IMDBRating, and MaxTheaters.



Left scatterplot shows relationship between opening gross and total gross; right shows relationship between IMDB rating and number of votes on IMDB. Red points did not breakeven; blue points did.

Appendix B: Logistic Regression Model

The target variable is "Breakeven". The success class=1, which means that a film did breakeven from box office sales.

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-5.30040979	0.70259899	0	*
Budget	-0.1557765	0.01487826	0	0.85575044
OpeningGross	0.39828286	0.04490587	0	1.4892652
IMDBRating	0.52095139	0.09765893	0.0000001	1.68362868
MaxTheaters	0.00137735	0.00016614	0	1.0013783
Summer Release	0.48275226	0.23492539	0.03988699	1.62052834
Christmas Release	1.22813332	0.3376188	0.00027516	3.41484904

Residual df	1001
Residual Dev.	593.456604
% Success in training data	33.03571429
# Iterations used	10
Multiple R-squared	0.53600442

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	247	86
0	26	649

Error Report			
Class	# Cases	# Errors	% Error
1	333	86	25.83
0	675	26	3.85
Overall	1008	112	11.11