Predicting Movie Revenue

Nikhil Apte, Mats Forssell, Anahita Sidhwa
December 16, 2011

Introduction

'Hollywood is the land of hunch and the wild guess' (Litman & Ahn).

Jack Valenti, president and CEO of the Motion Picture Association of America, once mentioned that '… 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' (Valenti, 1978).

Movie revenue depends on multiple factors such as cast, budget, film critic review, MPAA rating, release year, etc. Because of these multiple factors there is no analytical formula for predicting how much revenue a movie will generate. However by analyzing revenues generated by previous movies, one can build a model which can help us predict the expected revenue for a movie. Such a prediction could be very useful for the movie studio which will be producing the movie so they can decide on expenses like artist compensations, advertising, promotions, etc. accordingly. Plus investors can predict an expected return-on-investment. Also, it will be useful for movie theaters to estimate the revenues they would generate from screening a particular movie.

Data

We obtained our dataset by extracting the relevant information from the Internet Movie Database (IMDb). At first we considered using the existing dataset from Maas & al, but this dataset did not contain all the information that we needed. For this reason we deceloped a python script to go through all the movies in the database and download two html pages for each movie: the main page ("imdb.com/title/'movie id'.html") and the box office page ("imdb.com/title/'movie id'/business.html"). The data we needed was spread among those two pages. Specifically, the movie genre (each movie belongs to one to three predefined genres) and release date were taken from the main page, while the total box office revenue, opening box office revenue, number of theaters on release and budget were taken from the box office page.

We restricted our search to movies released after January 1st, 1990 and only included those having total box office revenue greater than $100,000 at first. Due to the high variability and incompleteness of the data for the low revenue movies, we then dropped movies whose revenue was smaller than $1,000,000. We also removed all movies for which some of the data was not available (typically either the budget or the number of theaters on release was sometimes missing). Our final dataset included 2510 movies. In the implementation of our different algorithms, we used hold-out cross validation to estimate testing error: we randomly selected 70% of the dataset to do the training and tested the algorithm on the remaining 30%.

To take into account the inflation rate, and the fact that for this reason more recent movies will find it easier to perform better, we adjusted the global box office revenue, the opening week-end revenue and the budget by dividing the values by the average price of a movie ticket on the year the movie was released and multiplying by the current average movie ticket price (this is somewhat of an approximation for

adjusting the budget, since the budget is depending on real inflation and not just movie ticket inflation, but we are taking these to be similar). For reference, the average movie ticket price has increased by more than 85% between 1990 and today.

Algorithms

Linear Regression

We first implemented a linear regression algorithm, with the first week-end revenue, budget and number of theaters on opening week as our parameters. The output of the algorithm is the global box office revenue.

The average testing error for this algorithm was around 110-130%.

The testing error plots show that the performance of the linear regression is very good for movies having high global revenue (and also in general high first week-end revenue, number of theaters and budget): we have error of less than 100% on these. For movies with low global revenue however, the testing error is sometimes very high, more than 2000% in cases. Lower revenue movies will require more sophisticated algorithms and more data to be analyzed better.

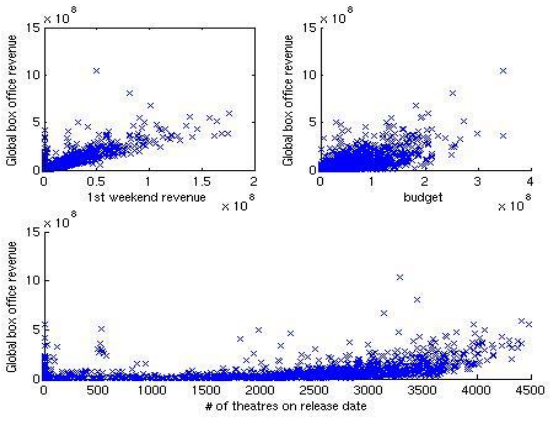

Figure 1: Training data – Global box office revenue vs1st weekend revenue, budget & no of theatres on release date

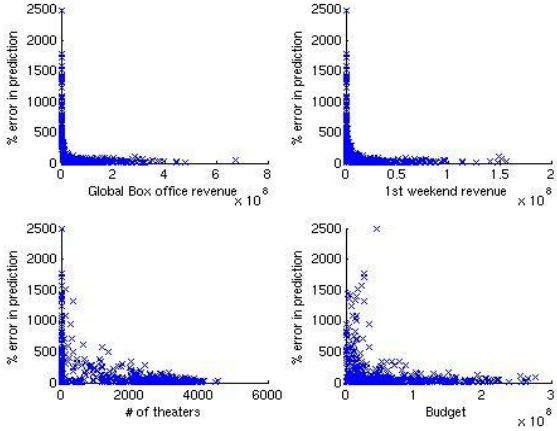

Figure 2: Prediction error for linear regression

From the plots of the training data, we also notice that while the global revenue is a strong linear function of the first week-end revenue, and also a linear function of the budget (with somewhat more noise), the global revenue is not really linear in the third parameter, the number of theaters. We can in fact roughly distinguish two regions from the plot of global revenue versus number of theaters: a 'low revenue' region and a 'high revenue' region. Therefore we decided to separate the data between these two regions and to run a linear regression and all subsequent algorithms separately on the two datasets.

K-means clustering

We separated the data using K-means clustering with two clusters, using only the number of theaters as our dataset for this algorithm.

Using this method, we find a testing error of around 30-40% for the cluster corresponding to movies released in a larger number of theaters, and an error of around 200-250% for the cluster corresponding to movies released in a low number of theaters

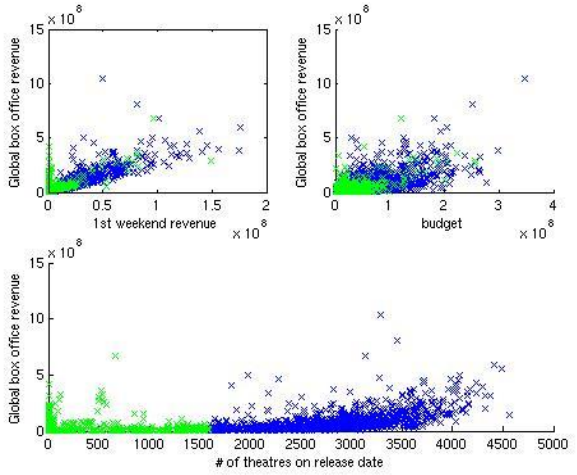

Figure 3: Training data after K-means clustering

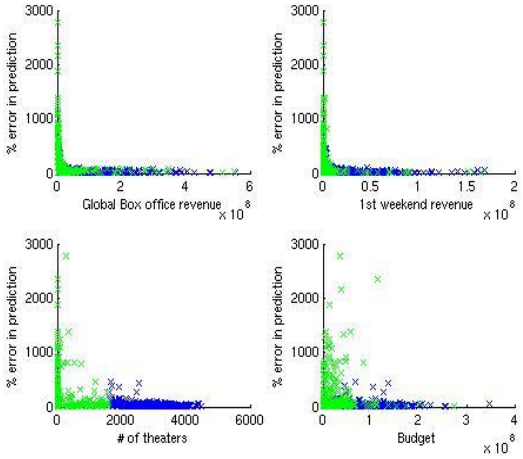

Figure 4: Prediction error after applying linear regression on the 2 clusters separately

Weighted linear regression

Although the global box office revenue has quite a linear variation with the first weekend revenue, its relation with budget and the number of screens is not exactly linear. To account for the nonlinearities, we decided to use a locally weighted linear regression model. In this model a separate hypothesis is made for each test sample with higher weights attached to the training data points which are closer to this particular test sample. The weighting factors are chosen according to

$$w^{(i)} = \exp\left(-\frac{x^{(i)} - x}{2\tau^2}\right)$$

The bandwidth parameter τ is chosen to minimize the mean testing error. The local weighting generally improves the accuracy of our prediction.

Polynomial Regression

Because the relation with budget and the number of screens is not exactly linear, we also used polynomial regression to account for the nonlinearities. We performed polynomial regression for order 2,3,and 4. From the results, we can see that overall, polynomial regression of order 3 provides the smallest mean error for total dataset, and high number of movie theaters.

Results

	All movies taken together			Movies released on high Number of theaters			Movies released on low number of theaters		
Algorithm	**Mean Error**	**Median Error**	**Std Dev**	**Mean Error**	**Median Error**	**Std Dev**	**Mean Error**	**Median Error**	**Std Dev**
2nd order Polynomial regression	63.1	34.0	125.6	29.6	25.9	22.2	191.3	69.4	377.6
3rd order Polynomial regression	61.8	45.1	103.1	28.4	24.8	24.2	99.4	71.7	178.3
4th order Polynomial regression	73.5	82.1	30.1	47.4	48.6	28.8	85.9	91.1	74.0
Linear regression	130.9	37.7	269.7	37.6	26.8	44.7	262.4	103.2	40.4

Table 1: Prediction errors for the different algorithms

From the results above, we observe that no specific method performed well for all three types of datasets. We observed that the higher the number of theaters, the more accurate the prediction. This is because the correlation between revenue, number of theaters, and budget is usually much better for movies that released on high number of theaters.

Genre Separation

Intuitively, it is easy to see that movies from different genres will perform differently in the theaters. For this reason we split our data according to genre, and ran our regression models for each data subset separately.

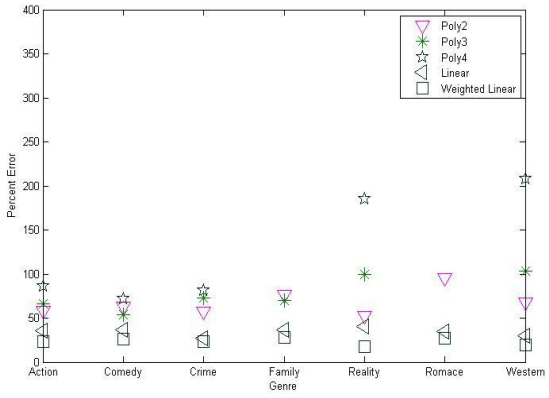

Figure 5: Mean error for each Genre (cluster with higher number of theaters)

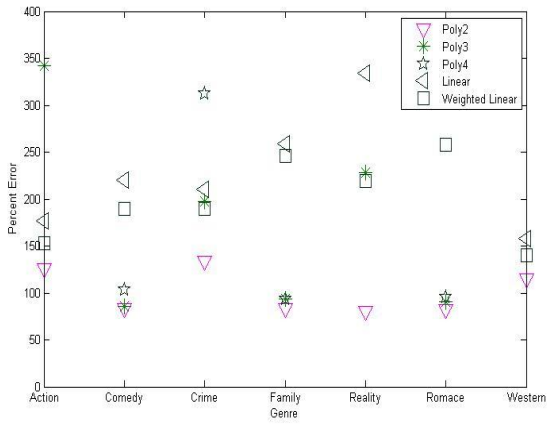

Figure 6: Mean error for each Genre (cluster with lower number of theaters)

Genre separation improved the accuracy for some subsets, especially for the movies which released on more number of theaters. However for subsets which did not have enough samples, the accuracy worsened. The third and fourth order polynomials over-fit the data for some genres, resulting in high testing errors. These outliers were excluded from the plots.

Conclusion

By combining k-means clustering to separate the dataset between 'low-release' and 'high-release' movies and separating the genres to take into account the diversity of movies, we managed to significantly improve the prediction accuracy using polynomial regression and weighted linear regression. In some cases it is possible to predict total revenue with accuracy better than 20%. However, some genres did not have enough samples in our dataset, making prediction for these genres very difficult. Moreover, even for the genres that we could train on, the 'low-release' movies cluster failed to perform well on average in every single case. It is worth noting that the information available for these smaller movies might not be fully accurate, adding some element of noise to the dataset.

In order to improve the predictions, attempting to separate the data along more than two clusters might be useful; however since this also reduces the number of available training examples for each subset, the performance increase will be limited. Also, the models can be improved by taking into account a lot of other factors, like the movie cast, the movie's release date (i.e. whether the movie released during holiday season, etc.), the movie's competition when it released, etc. Another step further could consist of considering the specificities of foreign movies, and of foreign markets.

Acknowledgement

We want to thank Professor Ng for the opportunity to research this project, and the CS229 CAs, in particular Andrew Maas, for their help and support. Chris Pouliot offered his advice on how to get started and ideas on what we should be aiming for. Chris Potts also provided us with his baseline python script for web crawling.

Works Cited

Litman, B. R., & Ahn, H. (n.d.). Predicting financial success of motion pictures.

Valenti, J. (1978). Motion Pictures and Their Impact on Society in the Year 2000, speech given at the Midwest Research Institute, Kansas City.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).*

The Internet Movie Database: http://imdb.com

National Association of Theater Owners: http://www.natoonline.org/statisticstickets.htm

Ng, Andrew, CS229 Lecture Notes