BOX OFFICE ANALYTICS

Date: April 20, 2016

Prepared By:
Akansha Khandelwal
Bharath C. Sharma
Chirag Mital
Parth Pareek

TABLE OF CONTENTS

I. Introduction 3
II. Data Preparation 3
III. Analysis 4
    a. Pre-production phase
    b. Pre-release phase
    c. Blockbuster prediction
IV. Limitation 7
V. Conclusion 7

I. Introduction

The movie industry today is seeing a multifold increase in investment with production budget records broken every year. However, only a fraction of these recover and surpass their investment in the box office. Production houses need to strategically invest their resources to ensure profitability and even avoid bankruptcy. The use of data from previous movie releases might aid production companies gauge the preferences of viewers with respect to cast, style or genre and determine if a particular idea is likely to be a success or not. In this project, we aim to create a support system for production houses to aid investment decisions in the pre-production as well as the post production, pre-release stages of the movie by determining factors of a movie's success during the production process. The goal of this project is to better quantify the elements that are significant determinants of financial success of a movie when making the decision to fund a motion picture production.

II. Data Preparation

We acquired data from three different sources for this project. The initial data for movies and user ratings was acquired from MovieLens. It comprised of three files for movie data, user demographic information and user ratings. For each movie, we had the release year, movie ID, titles identical to titles provided by IMDB and genres. Genres were pipe separated and were selected from the following: Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War and Western. To make our analysis easier, we grouped similar genres.

Genre 1 - Animation, Children, Fantasy, Musical
Genre 2 - Action, SciFi, War, Western
Genre 3 - Crime, Film-Noir, Horror, Mystery, Thriller
Genre 4 - Comedy
Genre 5 - Documentary
Genre 6 - Drama, Romance

We had 1 million ratings by 6040 users for 3952 movies. Ratings were made on a 5 star scale and every user had rated at least 20 movies. We also had demographic information such as age and gender for all users. Age is chosen from the following ranges: Under 18, 18-24, 25-34, 35-44, 45-49, 50-55 and 56+.

We then scraped movie metadata from IMDB using Python. This data comprised of features such as Actors, Director, Production House for each movie. We identified a list of top 30 male actors and assigned each movie a binary value of 0 or 1 if the male actors of that movie were considered top actors or not. We used the same logic for top 30 female actors, top 30 directors and top 10 production houses. We naively assume that every top male or female actor, director and production house in this list is equally important. We also identify if a particular movie is a sequel or not and assign it a binary value of 0 or 1 based on this.

We then scraped data for Gross Domestic Revenue and Budget from Box Office Mojo using Python. We tried to fill missing values for these features from IMDB as well and converted entire data into USD.

After acquiring data from these sources, we collated the entire data and aggregated ratings for all user demographics. We filled missing values for ratings using K-Nearest Neighbors with k=5. After removing movies with missing values for revenue and budget, our final data set comprised of 1473 movies with 29 covariates.

III. Analysis

a. Pre- Production Phase Predictions

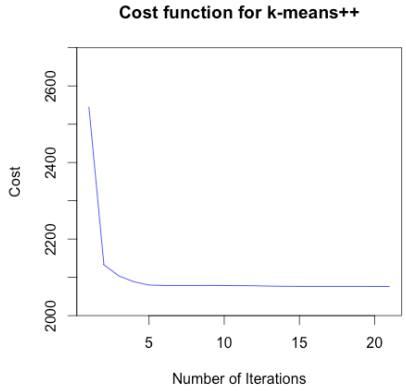
Unsupervised learning to get value from movie meta data. A snippet of the dataset is shown below.

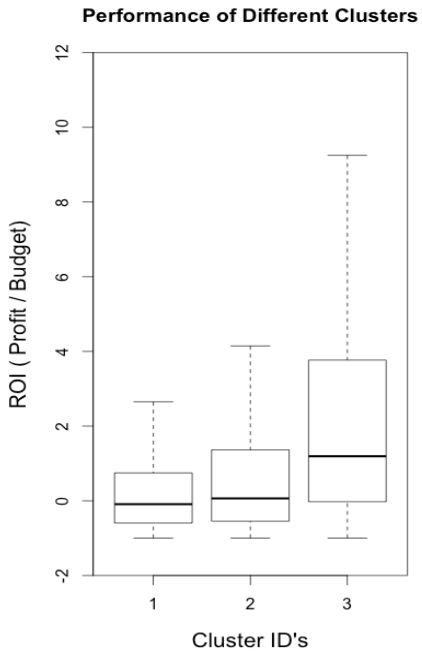
Movie	Year	Sequel	Budget	Top Actors	Top Actresses	Top Directors	Top Production	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6
Armageddon	1998	0	1.4e+08	1	0	0	0	0	1	1	0	0	0
Lethal Weapon 4	1998	1	1.4e+08	0	0	1	1	0	1	1	1	0	1
Small Soldiers	1998	0	4.0e+07	0	0	0	1	1	1	0	0	0	0
Pi	1998	0	6.0e+04	0	0	0	0	0	1	1	0	0	0
Chariots of Fire	1981	0	5.5e+06	0	0	0	1	0	0	0	0	0	1
Terms of Endearment	1983	0	8.0e+06	1	0	0	1	0	0	0	1	0	1

K Means ++ Clustering

We use data on 1293 Movies to perform clustering. We use a heuristic to find the initial centers in the algorithm. This is then followed by the standard Lloyd's algorithm to find a locally optimal clustering. This method outperforms the standard K Means algorithm, both in terms of the time complexity and the final local optimum that is reached. This is highlighted by the steep decline in the figure which shows the value of the cost function vs time.

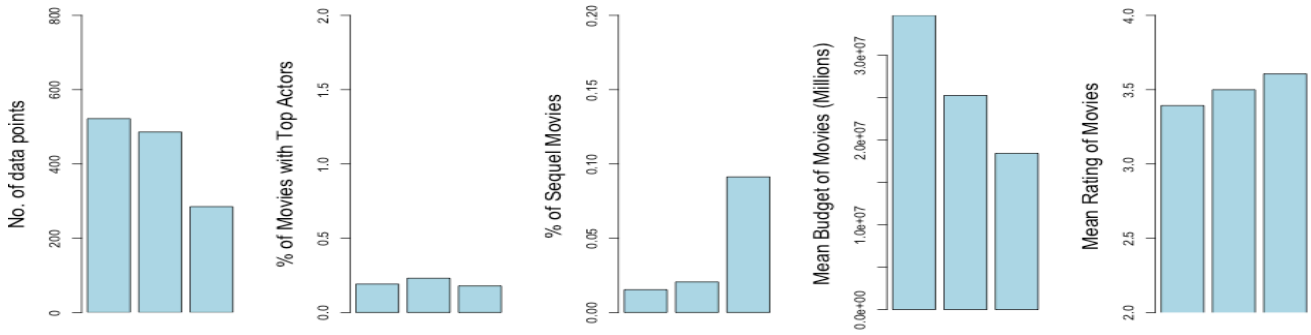
We used the Euclidean Distance as the similarity measure and chose k = 3 clusters for this problem. We also assigned small weights to “Year” and “Sequel” to emphasize their importance with respect to the other covariates.





The value proposition of clustering can be seen by analyzing the difference in the performance of Movies across clusters. The figure on the right shows a stark contrast between the interquartile ranges of the ROI's of movies in each clusters. This can be used to get an expected range of ROI just based on the meta data. Given that this can be done even before a movie goes to production is a definite value add for production houses.

Furthermore, the variation of parameters across the 3 clusters provide some interpretability to the model. We see that % of Sequels are much higher in cluster 3 and that mean Budgets and Ratings show a weak correlation with ROI's.



Clustering can be a powerful tool especially when the number of data points n >> p (number of covariates). Given more data, the differences between the clusters will be more pronounced and more significant statistically.

b. Pre-Release phase revenue prediction

To aid decision making in the post-production pre–release stage we use supervised learning techniques to predict the revenue of movies based on budgets, metadata and focus group ratings. To bring the data to the same scale for prediction, we use log transformed revenues and budgets. The log transformed revenue is regressed against all other variables and Mean Squared Error (MSE) is used as the criteria to determine the performance of the model. The model is trained on a training set of 75% and tested on the remaining 25% of the data and the MSEs are reported. (As all MSE is of the log transformed data, the true MSE in Revenue prediction is e^{MSE})

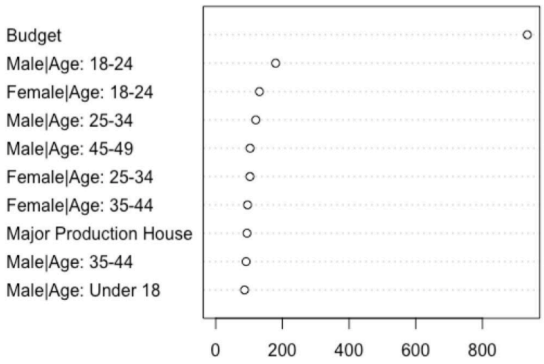
First, we performed Linear Regression on all variables performs fairly poorly, giving a MSE of 2.01. The log of revenue is predicted as a linear function of 27 covariates including log of budget, movie metadata and focus group ratings.

Secondly, we performed forward stepwise subset selection to determine the best variables that predict the revenue using adjusted R^2 as the performance criterion. The MSE obtained is 1.98 and is only marginally better than the full model linear regression. Out of 27 covariates, 20 were selected.

Further, we performed lasso regression to reaffirm our subset selection. With 5-fold cross validation, we obtained a shrinkage parameter of 0.01. An MSE of 1.99 was obtained with similar covariates as subset selection being drawn and similar adjusted R^2. Ridge regression (lambda = 0.1) also produced unconvincing results (MSE = 1.99). This

suggests that a linear model does not describe the data and other non linear prediction techniques would have to be explored.

Trees and Random Forests were used to predict revenues giving an MSE of 2.07 and 1.56 respectively. The random forest also provided a list of most important parameters which is provided below along with a summary of the results of all the techniques.



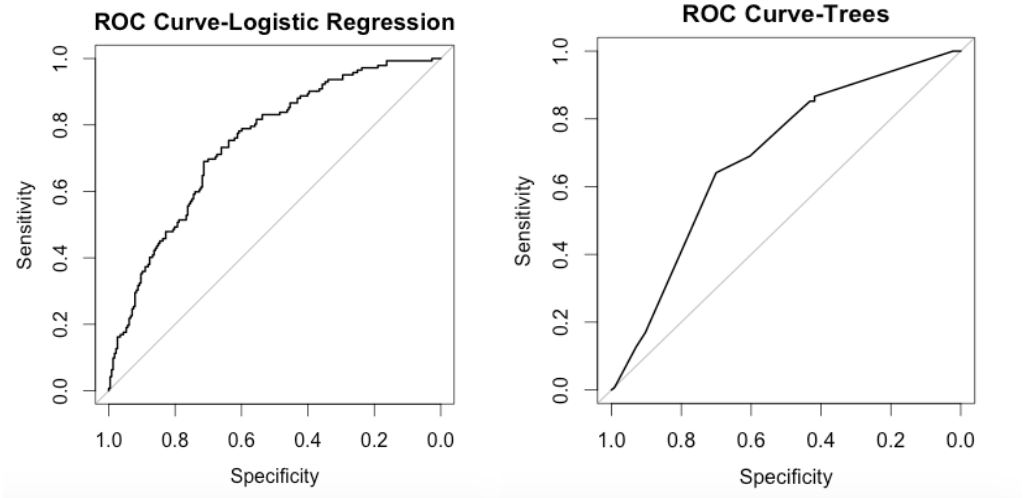
Variable importance based on Random Forest

Method	MSE
Linear Regression	2.01
Subset Selection	1.98
Ridge	1.98
Lasso	1.99
Trees	2.07
Random Forest	1.56

c. Blockbuster Prediction

In the previous part, we inferred that linear models are not the best choice when working with user demographic information. Therefore, we decided to use classification to decide whether a movie will be a blockbuster or not. Blockbuster is defined as any movie that generates a return on investment of more than 2 times the budget. Here, we focus on obtaining a good conservative result, i.e, reducing the False Positive Rates (FPR) to reduce misclassification of non-blockbusters as blockbusters.

First, we run a logistic regression to predict if a movie will be a blockbuster or not and we get a False Positive rate of 21.93%.



truth \ predict	0	1
0	178	50
1	59	82

Logistic Regression

truth \ predict	0	1
0	172	56
1	61	80

Trees

truth \ predict	0	1
0	188	40
1	56	85

Random Forest

Confusion Matrices

To test the models, we use ROC plots to check the performance against completely random allocations. The area under the ROC curve for logistic regression is 0.73.

We use a similar approach for classification with trees and get a false positive rate of 24.56%, which is slightly worse than logistic regression. Area under the ROC curve is 0.69.

Lastly, we classify using Random Forests and obtain a False Positive rate of 17.10%. We conclude that random forests' model gives the best performance.

IV. Limitations

There were a few limitations to our project. There are possible biases that exist as far as sampling goes. There was limited availability of revenue and budget data that introduces bias. We had to remove 2000 movies from our analysis due to this reason. Another limitation was that we included only box office collection as a measure of revenue. For many movies, revenue is also generated from sales and rentals of DVDs, sale of movie merchandise, etc. Lastly, we didn't have ratings for certain user demographics which had to be filled using KNN heuristic. This added a little bias to our analysis.

V. Conclusion

The results of our project suggest that, based on the limited information, accurate predictions cannot be made. Predictions based on social media data and sentiment analysis have shown better results and we recommend future work to consider this avenue.