Predicting Movie Box Office Gross

Joanna Khek Cuina U1440938K
Koh Ngiap Seng U1440549H
Ong Wee Kiat U1440330L

November 19, 2017

1 Introduction

A forecast in 2016 predicted the global box office revenue to increase from about 38 billion U.S. dollars to nearly 50 billion U.S. dollars in 2020 [1]. With the increase in money invested in the makings of a movie, it has become paramount that movies are successful to justify these large undertakings. What makes a movie successful? The presence of star actors did not necessarily do so. Rush Hour 3, a movie co-starring Jackie Chan and Chris Tucker went into history as one of the Top 100 Box Office losses [2]. Neither did movie ratings too. Transformers: Age of Extinction were a huge commercial success, but ended up receiving poor ratings [3]. While many factors play a role in the success or failure of a movie, it remains unclear how they interact, or how much of a role they play in deciding the success of a movie. Hence, this project aims to predict the movie box office gross based on pre-movie release factors through the use of various techniques in Data Mining.

2 Data Acquisition and Cleaning

The dataset selected is the TMDB 5000 Movies dataset [4] which was made up of 4803 movies released between 1916 to 2017.

Movies used in this project were restricted to those released between 1990 to 2017 as varying tastes in movies across different eras may lead to computational difficulties. In addition, only movies that had a minimum budget of $100,000 were considered to reduce the number of independent movies which have limited information available. In addition, as mentioned in the introduction, only predictors that are available before its debut were considered. As such, these variables were excluded from the dataset: “popularity”, “vote_count” and “vote_average”.

In order to account for inflation over the years, and the fact that for this reason more recent movies will find it easier to perform better, the revenue and budget was adjusted based on the Consumer Price Index (CPI) [5]. By setting the most recent year 2017 as the reference point, the CPI of the newer year was divided by the older year, followed by multiplying the unadjusted data by the ratio obtained. The inflation-adjusted data is shown in Figure 1 below.

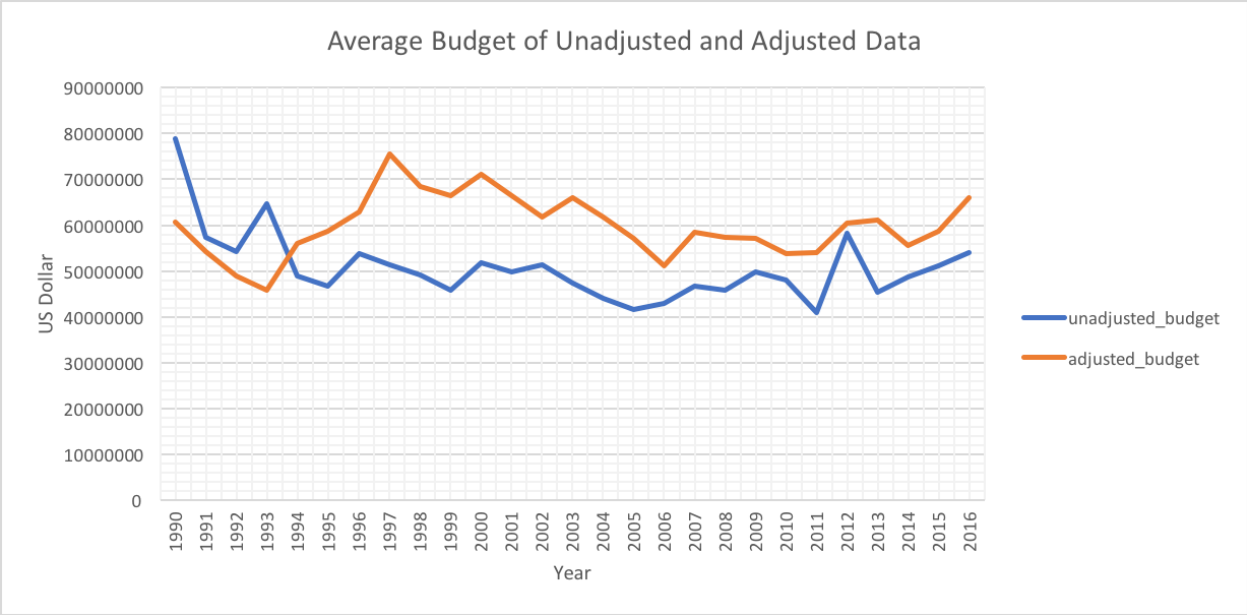

Figure 1: Average Budget of Unadjusted and Adjusted Data

Due to the huge number of actors, producers, directors and production companies, these variables were replaced with binary variables named "best_actor", "best_picture", "best_director" and "top_companies", which were grouped based on:

$$\text{best\_actor} = \begin{cases} 1, & \text{if actor awarded the best actor award at The Oscars [6],} \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{best\_picture} = \begin{cases} 1, & \text{if producer awarded the best picture award at The Oscars [6],} \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{best\_director} = \begin{cases} 1, & \text{if director awarded best director award at The Oscars [6],} \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{top\_companies} = \begin{cases} 1, & \text{if within top 20 production companies list [7],} \\ 0, & \text{otherwise.} \end{cases}$$

For our genres, we converted them into binary variables.

$$\text{genre} = \begin{cases} 1, & \text{if the movie is of the genre,} \\ 0, & \text{otherwise.} \end{cases}$$

The impact of the holiday season was also a consideration taken into account in this project (which was set to be month of November, December and January) and may thus play an important role

in our overall prediction. Hence, another variable "holiday_season" was also introduced, which indicated if the movie was released during the holiday season.

$$\text{holiday\_season} = \begin{cases} 1, & \text{if the movie is in release\_month 11, 12 or 1,} \\ 0, & \text{otherwise.} \end{cases}$$

3 Features

The final dataset was reduced to 2385 movies. Below are the features utilized in the prediction.

Feature Categories	Features	Converted Form	Examples
Qualitative	Genres	binary	Action, Adventure, Romance
	Plot	binary	In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but becomes torn between following orders and protecting an alien civilization.
	Production Companies	binary	Twentieth Century Fox Film Corporation
	Producers	binary	James Cameron
	Directors	binary	Steven Spielberg
	Actors	binary	Leonardo DiCaprio
Quantitative	Budget	numerical	268758000
	Revenue	numerical	3161552409
	Runtime	numerical	120
	Released Year	numerical	2010
	Released Month	binary	January

Table 1: Features Utilized in Our Prediction

4 Preparation

The data was separated into training and test data randomly. Namely, 70% training data and 30% test data. Training data was then normalized to reduce the bias of predictions arising from large values. Methods that required cross validation had normalization done on the pseudo training set instead. For normalization, binary variables were divided by the standard deviation as they were best interpreted if they were either 0 or not. Non-binary variables were normalized by the mean and divided by the standard deviation.

5 Methods

5.1 Multiple Linear Regression

Multiple Linear Regression was the initial approach adopted due to its ease of implementation as well as relatively few assumptions. The response variable was the revenue whereas the input

variables consisted of all the other features. An adjusted R-squared value of 0.4723 was obtained which was not as good as expected.

The possible existence of interaction between factors was also considered. An analysis of the correlation matrix suggested a possible relationship between "budget" and "Adventure", which was incorporated into the model, in turn leading to a slight improvement in the adjusted R-squared value and test error. The results are summarized below:

	Without interaction	**With interaction**
Adjusted R-squared	47.23%	47.64%
Test Error	37.27%	36.17%

Table 2: Test Error for Multiple Linear Regression

5.2 Logistic Regression

Next, Logistic Regression was performed as it was most comparable with the Multiple Linear Regression which also utilized linear weights. The data was grouped into different bins according to their respective revenues (in millions) and each bin contained approximately same number of data to ensure proportionality. The misclassification rate obtained was 24.4%

Bins	**1**	**2**	**3**	**4**	**5**
Revenue	0.1M - 10.86M	10.86M - 24.4M	24.4M - 41.4M	41.4M - 64.2M	64.2M - 91.5M
Bins	**6**	**7**	**8**	**9**	**10**
Revenue	91.5M - 128.4M	128.4M - 180.3M	180.3M - 260.8M	260.8M - 453.9M	$>$ 453.9M

Table 3: Bins Categorization for Revenue (in millions)

5.3 Regularization and Shrinkage Methods

Although the application of Logistic Regression led to a lower misclassification rate, there was still the possibility that other models could yield better results. Hence, Regularization Methods was implemented on the data next.

	Forward Stepwise (R^2)	**Forward Stepwise (AIC)**	**Forward Stepwise (BIC)**	**Ridge Regression**	**LASSO**
Test Error	37.28%	37.16%	36.81%	36.15%	36.59%

Table 4: Test Error for Regularization Methods

From the results, Ridge Regression yielded the smallest test error which suggested that further removal of more variables was not ideal.

5.4 Artificial Neural Networks (ANN)

For Artificial Neural Network, the best neuron size was obtained by repeating the function several times. A size of 11 was found to give the best prediction and only 1 hidden layer was considered due to time constraints coupled by the fact that the dataset contained too many predictor variables. A function that performed repetitions to determine the weights required to minimize the cost was also created. The misclassification rate was 10%, which was better compared to other models.

5.5 Regression Tree

A very large tree was grown initially and cross validation was performed on it. Regrettably, the number of terminal nodes obtained after cross validation was the same as the full tree. The conclusion drawn was that the full tree was the best tree and simpler trees were worse off.

Various aggregating methods were also implemented in an attempt to improve our tree. The aggregating methods performed were Bagging, Random Forest and Boosting. It was worth noting that Random Forest gave the smallest test error in comparison to Boosting and Bagging. Interestingly, some significant variables were: "budget", "production_company", "drama", "animation", "runtime" and "family".

	Full Tree	**Pruned Tree**	**Bagging**	**Random Forest**	**Boosting**
Test Error	48.15%	48.15%	37.87%	31.85%	42.84%

Table 5: Test Error for Regression Tree

5.6 Classification Tree

Utilization of the Classification Tree led to a similar issue being encountered in the usage of the Regression Tree. The full grown tree yielded the same number of terminal nodes despite cross validation being applied. By performing Random Forest, a minor improvement in the misclassification rate was obtained.

	Full Tree	**Pruned Tree**	**Random Forest**
Misclassification Rate	76.96%	76.96%	75.28%

Table 6: Misclassification Rate for Classification Tree

5.7 Naïve Bayesian Classifier

In using the Naïve Bayesian Classifier, independence between attributes was an assumption even though in reality, they may be dependent in some way. Since there was a possibility that simpler models could yield the best result on some occasions, Naïve Bayesian Classifier was implemented.

The Naïve Bayesian Classifier is primarily based on the following formula:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

For the analysis, A was set to be the response variable - "bins", whereas B consisted of 28 predictor variables. The output would then represent the bin with the highest probability given the relevant inputs. Laplace smoothing was not required in the analysis since the variables involved were all binary, hence implementation of Laplace smoothing would not have any significance. A misclassification rate of 53.6% was obtained.

5.8 Text Mining

A possibility of movie plots playing a significant role in affecting revenue generated was also explored to see if the more popular plots were indeed significant in generating the revenues.

To begin, Text Mining was implemented to obtain words that appeared frequently. Stop words were also introduced - which were essentially words that were insignificant to the plot but appeared too frequently. A Multiple Linear Regression was then conducted upon removal of those words.

The results are summarized below:

	Without Words	**With Words**
Adjusted R-squared	47.23%	47.99%
Test Error	51.88%	38.42%

Table 7: Test Error for Text Mining

Evidently, both the adjusted R-squared value and the test error improved with the addition of those words. Interestingly, some of the more significant words were "dead", "help", "home" and "save".

5.9 Support Vector Machine (SVM)

A multinomial Support Vector Machine model was also implemented to classify movies into bins which were partitioned previously.

The "one-versus-one" strategy was used to decompose the multi-class problem into several two-class sub problems, where each binary problem was solved by using the standard SVM. Training models were then tested using Linear, Polynomial and Gaussian kernels. In addition, ten-fold cross validation was performed to compare SVMs with the same kernel using a range of values of the cost parameters.

The misclassification rate and the best parameters selected from the cross validation for each kernel is shown in Table 8 below.

	SVM Linear Kernel	SVM Polynomial Kernel	SVM Gaussian Kernel
Misclassification Rate	8.38%	67.18%	32.40%
Parameters	Cost = 100	Cost = 100, Degree = 3	Cost = 1000, Gamma = 0.01

Table 8: Misclassification Rate for Different Kernels of SVM

6 Conclusion and Discussion

Overall, the test error and misclassification rate improved through the application of more complex models such as Support Vector Machine and Artificial Neural Network. Regression Tree performed the worst among all the other models, which was expected and it could probably be due to the non-robustness of the regression tree - a small change in the data could result in a large change in the final estimated tree.

On the whole, it was also noteworthy that the features used were insufficient to provide a very convincing prediction of the revenue. Perhaps consideration of the MPAA (Motion Picture Association of America) film rating system, which was not available in the dataset that was used, could be considered in future works. Furthermore, the prediction could probably improve even further had the relationship between movie sequels been taken into account. Nonetheless, most of the methods produced results which were expected, indicating that our project could be a platform for other studies to build on in the future.

References

[1] Steve Fuller. *Film and Movie Industry - Statistics & Facts*. 2016. URL: https://www.statista.com/topics/964/film/.

[2] LLC Nash Information Services. *Movie Budget and Financial Performance Records*. URL: http://www.the-numbers.com/movie/budgets/.

[3] Colin Yeo. *Will the superhero films ever end? The business of blockbuster movie franchises*. June 2017. URL: http://theconversation.com/will-the-superhero-films-ever-end-the-business-of-blockbuster-movie-franchises-78834.

[4] Kaggle Inc. *TMDB 5000 Movie Dataset*. Sept. 2017. URL: https://www.kaggle.com/tmdb/tmdb-movie-metadata/.

[5] Federal Reserve Bank of St. Louis FRED. *U.S. Bureau of Labour Statistics, Consumer Price Index for All Urban Consumers: All Items [CPIAUCSL]*. Sept. 2017. URL: https://fred.stlouisfed.org/series/CPIAUCSL/.

[6] Academy of Motion Picture Arts and Sciences. *The Official Academy Awards Database*. Feb. 2017. URL: http://awardsdatabase.oscars.org/.

[7] LLC Nash Information Services. *Movie Production Companies*. URL: http://www.the-numbers.com/movies/production-companies/.