Communications for Statistical Applications and Methods
2013, Vol. 20, No. 4, 301–309

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.4.301

Predicting Gross Box Office Revenue for Domestic Films

Jongwoo Song[1,a], Suji Han[a]

[a]Department of Statistics, Ewha Womans University

Abstract

This paper predicts gross box office revenue for domestic films using the Korean film data from 2008–2011. We use three regression methods, Linear Regression, Random Forest and Gradient Boosting to predict the gross box office revenue. We only consider domestic films with a revenue size of at least KRW 500 million; relevant explanatory variables are chosen by data visualization and variable selection techniques. The key idea of analyzing this data is to construct the meaningful explanatory variables from the data sources available to the public. Some variables must be categorized to conduct more effective analysis and clustering methods are applied to achieve this task. We choose the best model based on performance in the test set and important explanatory variables are discussed.

Keywords: Box office, regression, linear model, random forest, gradient boosting.

1. Introduction

The year 2012 has been a special year for the Korean film industry. This year, there were a number of successful domestic movies that attracted millions of moviegoers. The Thieves (released in late July) has set a box office record with over 13 million viewers on that places it first in all-time box office history. In addition, Masquerade (released in mid-September) had a huge box-office success recording the fourth largest total attendance with over 11 million tickets sold. The recently released, a Werewolf Boy, is rapidly attracting record numbers of moviegoers (over six million as of November, 2012). In the first half of 2012, the biggest blockbuster film was the Hollywood movie, Avengers; however, seven domestic movies were among the top ten blockbusters leading the box-office in the first half of 2012. The Korea Creative Content Agency (KOCCA) has forecast that the gross revenue of the domestic movie industry is expected to grow by 4.4% relative to the first half of 2012; in addition, it states that the top four distributors accounted for 95.72% of the total market share of the audience, further pushing the polarization of major distributors and medium-sized distributors. The top four distributors include CJ E&M, Lotte Entertainment (Lotte Ent.), Next Entertainment World (NEW) and Showbox/Mediaplex Inc. (Showbox). From 2008 to 2012, 187 movies out of 528 movies belonged to the top distributors, with their gross revenue accounting for approximately 83.8% of the total gross revenue; however, there are various determinants of the box office success besides the distributors. It costs money to make a movie; therefore, a studio and investors want to predict the gross revenue that a movie will bring before a movie opens in theaters. There were several studies to predict box office revenue using various regression methods. Sharda and Delen (2006) applied neural networks and Neelamegham and Chintagunta (1999) used a Bayesian model to predict box

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No. 2013-012817).

[1] Corresponding author: Associate Professor, Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 120-750, Korea. E-mail: josong@ewha.ac.kr

office revenue. We can also find that there are other papers for box office predictions (Sawhney and Eliashberg, 1996; Terry *et al.*, 2003; Vany and Walls, 1996, 1999); however, these papers mostly used data from American box offices. We endeavor in this study to find a model for domestic movies.

The determinants of box office success and large size revenue may include star actors, genres, release date and others. In this paper, these factors and many others will be explored and considered to predict the gross revenue of movies as accurately as possible. Foreign movies (especially Hollywood movies) have great popularity in Korea; however, only domestic movies will be considered for analysis in this paper. The prediction model will be constructed based on Korea box-office data from 2008 to 2011. Three regression models (that include Linear Model, Random Forest and Gradient Boosting) will be considered and briefly introduced in Chapter 3. Using described methods, the domestic film data of four years from 2008 to 2011 will be analyzed. The data set will be described in detail along with the variable selection and categorization of some variables. In the last part of this paper, the performance of each model will be evaluated followed by a conclusion. Statistical software R (R Core Team, 2013)was used to analyze the given data; however, R package 'randomForest' was used for the Random Forest (Breiman, 2001) method and the package 'gbm' was used for Gradient Boosting (Friedman, 1999a, 1999b).

2. Data

2.1. Data collection

Among approximately 505 domestic films were released during 2008 and 2011; subsequently, this paper explores 206 domestic movies with gross box office revenues over KRW 500 million. The Korea Film Council (KFC) provides a Korea Box Office Information System (KOBIS) that contains various useful statistical data; subsequently, most data used in this paper were collected from KOBIS (www.kobis.co.kr). In addition to the factors provided by KOBIS, other possible explanatory variables were considered such as vacation seasons, opening season and temperature, star actors, and directors. The data were collected through various sources. Among all available data sources, a portal website with a movie information category of Naver (www.naver.com) was used most frequently for an additional data search. Websites used to obtain relative information include: www.kobis.co.kr, and www.kofic.co.kr, www.naver.com/weather, www.naver.com/movie.

2.2. Data preparation

We consider several explanatory variables in this section. Our goal is to find relevant variables to predict box office revenues as accurately as possible. We consider the following variables for the important explanatory variables.

2.2.1. Director

In 2012, National Geographic Korea introduced three directors in a three-episode project. What these directors shared in common was that they all had produced Korean blockbusters. They were so called blockbuster making directors known to the audience. When these directors put their movies on screen, many people expect these movies to be hits and thus these movies become blockbusters once again attracting a number of audiences. Therefore, the name value of the directors does matter. What we want to understand is how much the name value of the director affects the box office performance. Among around 500 films opened on screens during 2008 and 2011, there were more than 400 directors. However, each individual director cannot be treated as an independent variable; subsequently, they need to be assigned to a certain number of categories before considered in a predictive model. We will explain how we categorize director variable according to their ratings in the next chapter.

2.2.2. Star value

The movie 'Sunny' was one of the biggest hit movies in 2011; in addition, the Thieves (2012) was one of the biggest hit movies as well as the all-time most viewed movie. These two movies both attracted a large number of moviegoers and a large box-office revenue; however, they are distinct when it comes to the main actors. Vany and Walls (1999) argue that the star power in Hollywood movies does not affect the opening day of the movie but extends the movies' run. In this paper, we want to understand whether the star value affect the box office success; subsequently, the number of main actors was added as a variable in constructing three regression models. What really matters is who the actors are not just number of main actors that play in a movie. Therefore, similar to the director factor, the top class actors and actresses should be defined and these stars should be categorized into some number of groups.

2.2.3. Genre

Different from other media content, the genres of films are firmly set from the beginning of the production. Certain genres draws public fancy as Korean Film Council (KFC) states that plot, genre and audience opinion affects the choices of moviegoers.

2.2.4. Sequel

Star Wars, Lord of the Rings and Harry Potter were loved by millions of people around the world. Just by observing the success of these movies, the sequel seems to be a factor that affects box-office revenue. In this paper, this factor will be included as a binary factor to build a suitable regression models to predict box-office revenue.

2.2.5. Release date

Release date and the box office performance should be somewhat related because the movie opened during the hottest summer ever recorded in 2012; therefore, the Thieves, resulted in a huge success. In a news article of OhmyNews, it states that November is when Chungmuro (a commonly used term to refer to the movie industry) considers the month with least audience attendance. The relationship between box office revenue and the release date can be observed in side by side box plots. Various variables can indicate the released date, and in this paper, binary variables that indicate holiday seasons and vacation time our added along with the release month, season and temperature related information such as minimum, maximum and median temperature as well as the temperature difference of the released month.

2.2.6. KMRB rating

Sharda and Delen (2006) state that a commonly used variable to forecast box office success by the Motion Picture Association of America (MPAA) rating assigned to movies. MPAA relates to only movies that play in American territories; however, the Korea Media Rating Board (KMBA) is responsible for rating films released in domestic territories as one of four movie ratings (All, 12 and under, 15 and under and 18 and under).

Table 1: Variables in the data set

Variable	Definition	Information
genre	genres of movies, categorical variable	1- Drama/Romance/Family, 2- Animation, 3- Action/Criminal/Thriller/Mystery, 4- Horror, 5- Comedy, 6- Documentary, 7- Others
rating	KMRB film-rating	1- all, 2- 12 and under, 3- 15 and under, 4- 18 and under
sequel	binary variable	1- sequel, 0- not
dir.rating	rating of directors	rating from 1–10 based on box office 11 for multiple directors
main actor	number of main actors	
AAAA. actors	number of AAAA actors	
AAA. actors	number of AAA actors	
AA. actors	number of AA actors	
A. actors	number of A actors	
distributor	major or independent	1- major, 0- independent
holiday	move release date	Lunar New Year's day, Chuseok, Valentine's day and Christmas day
vacation	vacation season	1- vacation, 0- not
season	season variable	1- spring, 2- summer, 3- fall, 4- winter
month	month of release date	
temperature	average temperature of month	min, max, med, diff
share	share of the domestic movies	percent
score	the score of gross box office	scores assigned by percentiles, 1–10

2.2.7. Market share

Most previously published literature to predict the success of movies focus on Hollywood films. Sharda and Delen (2006) show the competition level as one of the important factors that affect box office success. By the level of competition, it refers to movies released at the same time or ones released prior but still playing in theaters. The Korean film industry differs from Hollywoods; therefore, this factor needs a bit different approach. The dominance of Hollywood movies around the world is clear; however, a unique case is shown in Korea where domestic films have strong presence and are highly preferred. In 1987, with the start of direct distribution of foreign films, Korean film industry had underwent a dark period until the big success of the domestic film 'Swiri' in 1999. Competition between movies in theaters is an important determinant for Hollywood box office success; in addition the competition between imported films and domestic films is an important factor of the success in the Korean film industry.

2.3. Data description

We have 206 movies with 19 explanatory variables from 2008 to 2011. There is no missing values in the data set. Table 1 shows the variables and their descriptions. Some variables are intuitive but others are not. We will explain how to construct some variables in the data set.

2.3.1. Scoring gross box office revenue

The distribution of gross box office revenue has very long right tail because a few movies have a large gross box office with relatively small values. Therefore, we will have very poor prediction power if we use raw box office revenue in the model. In order to increase the prediction accuracy, the box office revenue was scored in ten digits (1 to 10) according to the corresponding percentiles. Table 2 shows the percentiles of the data set obtained for this paper from 2008 to 2011.

Table 2: Percentile of gross box office revenue

Percentile	Revenue (KRW)
10	832,352,000
20	1,411,812,000
30	2,531,239,084
40	3,795,149,333
50	5,138,108,000
60	7,591,008,500
70	10,901,292,472
80	15,659,201,000
90	22,727,852,000
100	81,034,853,333

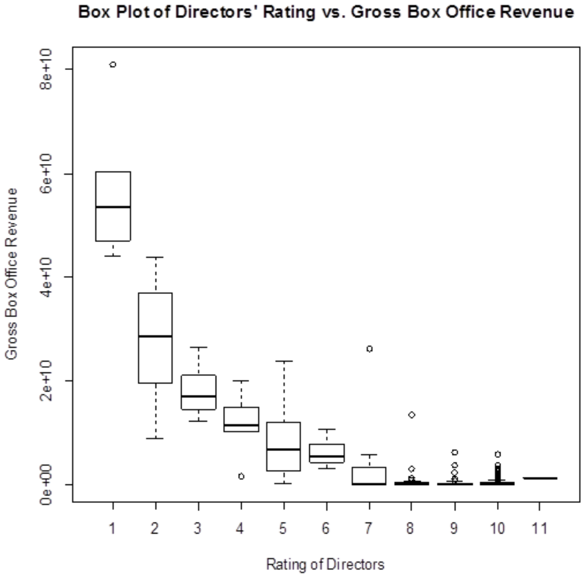

Figure 1: *Box plots of box office revenue based on the Directors rating*

2.3.2. Directors' rating

It is necessary to categorize all directors into some number of categories. We group directors based on previous box office performances; therefore, we did clustering analysis based on the directors' previous maximum box office revenue. We have used R and the clustering package 'mclust' to do clustering for this data. The package 'mclust' uses a Gaussian mixture model and can find the optimal number of clusters based on Bayesian Information Criterion (BIC). The optimum number of categories returned was ten from the mclust package. We have one more category for director's rating which is the movies for the multiple directors. Therefore, for each movie there is the director's rating variable with 11 groups. We put labels such that the first group of directors with the group number 1 represent the top dollar making movie makers. The name of this variable is 'dir.rating' in the dataset. As we can see from 1 there are 11 groups of directors based on their box office performance. We can see that first 4 groups of directors' box office revenue decrease significantly but the difference becomes smaller after group 5.

Table 3: Top rated directors and actors

Rate	Directors	Actors
1	Kang Hyngchul, Kim Yongwha	Won Bin, Kang Dongwon
1	Youn Jaekyun, Lee Jungbum	Kang Yewon, Knag Jiwhan
2	Kang Woosuk, Kim Jiwoon	Kyun Miri, Ko Doosim
2	Na Hongjin, Jang Hoon	Ko Su, Ko Changsuk, Kwon Sangwoo

2.3.3. Actors' rating

Main actors and actress of movies were categorized similar to the method used for directors. What was different was that instead of considering all main actors and actress, only those who had acted in movies with a gross revenue of more than KRW 1 billion were considered. The rest was very similar to the process of making directors' rating variable. We have 4 groups of actors' rating variables and we put labels such as 'AAAA actors', 'AAA actors', 'AA actors', and 'A actors' based on box office performances. We have included the top rated actors and directors in Table 3.

3. Analysis

In this chapter, we briefly explain three regression models and describe the results.

3.1. Models

We consider three regression models, linear model, Random Forests model, and gradient boosting model.

3.1.1. Linear model

A linear model is very popular for regression analysis. It is still most commonly used method because it is easy to fit and easy to interpret the results. It shows good performance in real data especially if the data is noisy. Since we have many predictors, we used stepwise procedure in R to find the model with the least Akaike Information Criterion (AIC) value.

3.1.2. Random Forests model

A tree-based method is very popular these days because it has several good properties. First, it is inconsequential if variables are categorical or numerical. Second, we can still fit the model even if the number of predictors is greater than the number of observations. Third, the fitting can be done very quickly.

There is a critical drawback for the single tree model. In general, it has less prediction power than other methods because of the lack of smoothness. One solution for this disadvantage is to bag trees. By bagging trees we can reduce the variance of estimator and improve the performance. Random Forest (Breiman, 2001) is a special case of bagging trees. It tries to reduce the correlation between trees by selecting the splitting variable from m randomly selected variables. By doing this, it can reduce the variance of the estimator better than naive bagging trees method. R package 'randomForest' can be used to analyze data using the Random Forest model.

3.1.3. Gradient Boosting model

Basically, boosting method uses several weak models and combine them to form the final model. It fits the model sequentially and the final model is the weighted average of the individual models. It is important to improve the model from the previous results. The gradient boosting model was introduced by Friedman (1999a, 1999b) and showed that it can compete with other top-level regression models in many cases. We can use R package ‘gbm’ for the gradient boosting model.

3.2. Important explanatory variables

In this section, we briefly explain the results of each model using whole data. We can find which variables are important in the model to predict the box office performance.

We found that the variables, genre, dir.rating, main.actors, AAA.actors, A.actors, distributor, vacation, season, month, and min.temp are significant in the linear model with stepwise procedure. The R-squared value is 0.6101, which is adequate for this kind of data. From Random Forest model, we found that dir.rating is most important variable followed by genre, A.actors and season. The gradient boosting method found that dir.rating is most influential variable followed by genre, season, and AAAA.actors. All three model found that dir.rating is the most important variable. The next important variables are genre, actor variables (star value) and season. These variables are reported important for all three models. Earlier in this paper, when the issue of sequels was discussed, we guessed that (different from Hollywood movies) domestic sequel movies would not help to succeed and that is what we get from the analysis results. The sequel variable is not a significant variable to predict box office revenue.

3.3. Performance in a test data

To find the optimal model, we partition the data ‘training set’ and ‘test set’. We randomly choose 70% of data for training (model fitting) and use the remaining 30% of data for testing (model evaluation). We have to do this procedure to find model to predict best for future data. The model that performs best in all data or training set can be an overfitting model that performs not as good as in the test data. We computed the mean squared error (MSE) for the test data and repeated this procedure for 1000 times. We used the number of trees is equal to 500 and $mtry = 6$ for the Random Forests model and the number of tree is equal to 500 and shrinkage factor is equal to 0.01 for the gradient boosting model. We can see the results from the following Table 4. We can also see the observed values versus fitted values plot in Figure 2. One thing to note is that random forest and gradient boosting models overestimate for small observed values and underestimate large observed values more than linear model does. It is because the fitted values of tree based method is the average of the observed values in the terminal node. We can see this trend in almost all cases when we use any tree based method.

4. Conclusion

Accurately predicting the box office performance of movies requires wide range of resources and information that makes the task is difficult to accomplish. Different from the Hollywood movie industry, data disclosure on the process of movie production is too limited for the domestic movie industry. This has been the largest barrier of this paper to forecast the box office revenue of Korean films. We believe our regression model can predict much better if we can get the information on the ‘budget’ like Hollywood films. Nonetheless, in a belief that the box office performance can still be predicted with given information, this paper explored three regression methods, Linear Regression, Random Forest and Gradient Boosting, to build suitable predictive models for domestic films using data from 2008 to 2011. In order to utilize the available data sources as much as possible, some variables such as directors related information and main star information were categorized to provide more explanatory information in regards to box office success. The result show that all three regression models

Table 4: MSE for test data

MSE	Linear model	Random Forest	Gradient boosting
Mean	4.556	4.062	3.946
SD	0.703	0.670	0.650

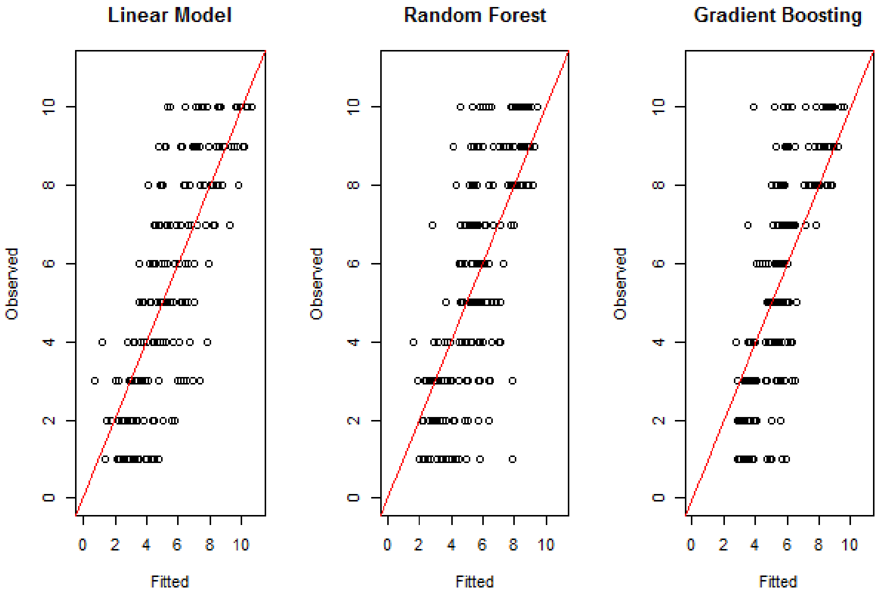

Figure 2: *Observed versus Fitted values plot for three models*

employed in this study can provide a moderate level of prediction for the box office success score of domestic films.

Among the three regression models, the gradient boosting model performs best in a test data (see Table 4). However, the performance difference is small and the linear model performs adequately. The R-squared value for the linear model is 0.588. The R-square value for random forest model is 0.537. The R-squared value for the gradient boosting is 0.604. Figure 2 shows that all three models perform similarly. We found that director's rating is the most important variable to predict the box office revenue. All three models we used for analysis to show the same result that director's rating is the most significant variable. Certain genres such as action movies and the existence of top class actors increase box office gross revenue.

References

Breiman, L. (2001). Random Forests, *Machine Learning*, **45**, 5–32.

Friedman, J. H. (1999a). Greedy Function Approximation: A Gradient Boosting Machine, Stanford University, http://www-stat.stanford.edu/jhf/ftp/trebst.pdf.

Friedman, J. H. (1999b). Stochastic Gradient Boosting, Standford University, http://www-stat.stanford.edu/jhf/ftp/stobst.pdf.

Neelamegham, R. and Chintagunta, P. (1999). A Bayesian model to forecast new product performance in domestic and international markets, *Marketing Science*, **18**, 115–136.

R Core Team (2013). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*.

Sawhney, M. S. and Eliashberg, J. (1996). A parsimonious model for forecasting gross Box-Office revenues of motion pictures, *Marketing Science*, **15**, 112–131.

Sharda, R. and Delen, D. (2006). Predicting box-office success of motion pictures with neural networks, *Expert Systems with Applications*, **30**, 243–254

Terry, N., Butler, M. and De'Armond, D. (2003). Determinants of the Box Office performance of motion pictures, *Proceedings of the Academy of Marketing Studies*, bf 8, 23–28.

Vany, A. D. and Walls, W. D. (1996). Bose-Einstein dynamics and adaptive contracting in the motion picture industry, *The Economic Journal*, **106**, 1493–1514.

Vany, A. D. and Walls, W. D. (1999). Uncertainty in the movie industry: Does star power reduce the terror of the Box Office?, *Journal of Cultural Economics*, **23**, 285–318.

Received April 16, 2013; Revised June 18, 2013; Accepted July 8, 2013