International Journal of u- and e- Service, Science and Technology
Vol.10, No.7 (2017), pp.25-34
http://dx.doi.org/10.14257/ijunesst.2017.10.7.03

A Study of Predict Sales Based on Random Forest Classification

Hyeon-Kyung Lee[1], Hong-Jae Lee[2], Jaewon Park[3], Jaehyun Choi[4]
and Jong-Bae Kim[5*]

[1]Graduate School of Software, Soongsil University, Seoul 156-743, Korea
[2]Department of IT Policy and Mgmt., Graduate School of Soongsil University, Seoul 156-743, Korea
[3,4,5]Professor, Graduate School of Software, Soongsil University, 156-743, Seoul, Korea*
[1]ketia89@naver.com, [2]hj1253@urpsys.com, [3]jwpark@ssu.ac.kr,
[4]jaehyun@ssu.ac.kr, [5]kjb123@ssu.ac.kr*

Abstract

The sales of movie industry have increased by 4.2% in 2015 compared to 2014 as reported by Korean Film Industry Council. This result can be attributed to the increase in the ticket price in addition to the expansion of the online market. Although South Korean's average annual movie consumption per capita is among the highest in the world, it is still difficult to estimate the probability of success for any given movie, and as such speculations come with high risks. Even among Holly Wood movies, only 2 or 3 out of 10 movies are successful, and there are many difficulties from development to release. Domestic movie industry also faces high risk, and the average profit from film investment in 2015 was at -7.2%, which shows the extreme difficulty of generating profit from investing in the movie industry. The attempts to minimize the risks by estimating the movie's success, such as attempting to estimate the number of audience based on quantitative data and deduction of variables, have been partially successful. However, due to the unforeseen effects of social phenomena, many of these predictions have also resulted in failures, which often inflicts in severe financial losses to the producers. This paper demonstrates the use of statistical approach to predict a movie's success, by analyzing the correlation between the total sales (dependent variable) and a number of potential influential factors (independent variables). In addition, the significance of each potential factor was quantified using Random Forest algorithm

***Keywords**: random Forest, Correlation Analysis, Predicting box-office values of movies, Predict analysis*

1. Introduction

The sales of movie industry have increased by 4.2% in 2015 compared to 2014 as reported by Korean Film Industry Council. This result can be attributed to the increase in the ticket price in addition to the expansion of the online market. Although South Korean's average annual movie consumption per capita is among the highest in the world, it is still difficult to estimate the probability of success for any given movie, and as such speculations come with high risks. Even among Holly Wood movies, only 2 or 3 out of 10 movies are successful, and there are many difficulties from development to release. The market size of film industry is surprisingly small compared with popular influence of movies. Domestic movie industry also faces high risk, and the average profit from film

[5*] Corresponding author. Tel. : +82-10-9027-3148.
Email address: kjb123@ssu.ac.kr(Jong-Bae Kim).

ISSN: 2005-4246 IJUNESST

investment in 2015 was at -7.2%, which shows the extreme difficulty of generating profit from investing in the movie industry[1].

With repeated full-scale growth in 2000s a synchronization with structure of Hollywood film industry is happening in Korea film industry. Media group represented by CJ E&M is having an increasing on overall industry with systematization not only in field of investment, distribution, and screening but also in related industry such as broadcasting, game, and internet. Distribution market of Korea film has changed in its structure from the type that production companies and small and medium size local distributors operated to the type that central distributor opens film at nationwide multiplexes at the same time. Recently, multiplex has become the type of theater itself occupying absolute majority of nationwide theaters. Investment source is also being expanded to video specialized investment unions, institutional investor, and financial circles and wide area opening has taken its place as an opening type of commercial film. Awareness of film industry practitioners has also greatly changed, and groups pursuing the rights of laborers are being organized and movement for treatment improvement is being realized.

There have been various attempts minimizing the risks of production cost and time in the movie industries. In particular, the attempts have been made to predict the box-office of movies accurately using several mathematical models and data mining method of econometrics. The attempts at predicting the number of the audience prior to the movie's release by extracting the variables based on quantitive data that has been partially successful. Based on these results, the movie companies have been selecting genre of movies that are popular in given seasons and maximizing the profit by effectively distributing the cost of production and marketing based on predicted numbers of audience.

However, in many cases, such predictions can be proven to be false, which inflicts great financial damage to the film producer. The unpredictability of the movie industry is caused by the unforeseen shifts in the overall entertainment industry, which reacts sensitively to social phenomena that is influenced by a large number of unknown factors. Furthermore, the viral factors over the internet, which are having increasing influences on the box office, are difficult to measure accurately. For example, while, 'Horror movie in summer' have been treated as the winning formula for box office success, this particular preconception has lost its power from several years ago. The previous studies regarding box-office have been conducted by the academia in the economics field, and these have only regarded factors such as, personal qualities of each directors and actors involved in the movie production, unique innate quality of the movie, or total capital investment for making the movie. This paper demonstrates the use of statistical approach to predict a movie's success, by analyzing the correlation between the total sales (dependent variable) and a number of potential influential factors (independent variables). In addition, the significance of each potential factor was quantified using Random Forest algorithm.

2. Related Study

2.1. Previous Research on the Success of Film

Byoung-Sun Kim(2009) analyzed the characteristics of movies based on the way it is released and the screening period, and their influence on the total number of audiences by categorizing the movie into types based on these factors. As a result, it is possible to perceive meaningful difference between the 'Wide release short period' type and 'Narrow release long period' type of movies. 'Wide release short period' types often indicate that these movies are fancy, showy, entertaining and has distinctive genres that can attract many audiences in a short period of time by opening in many theaters. On the other hand, 'Narrow release long period' movies have little entertaining qualities and aren't expected to attract many audiences simultaneously, and therefore they are screened in a small number of theaters for a longer period of time. In the case of 'Wide release short period' types, dramatic decrease in screening theaters and number of audiences have been seen in the early stage of release, whereas 'Narrow release long period' types show a smoother increase in both factors during the early stage of opening. In addition, for the case of 'Narrow release long period' movies, the factors that are applied to existing box-office research don't seem to have any effect, which means that new factors have to be regarded[2].

Sun Ju Kwon(2014) predicted the box office results by analyzing the number of articles from the media and NAVER movie ratings data. After simplifying the model by assuming the typically cited factors in box-office research such as the genre, actors and actresses, directors, seasons *etc.* are reflected on the number of screening theaters, the box-office result in each time period was analyzed for variables that have potential influence on the success. As most of the sales occur within the first 3 weeks of opening date, the analysis was done by dividing the period as pre-release, opening week, week 2, and week 3. Analysis showed that relevant variables change dependent on each time period. For pre-release and the opening week, the number of screening theaters and the number of audience were relevant variables in case of Korean movies. In case of foreign movies, the number of screening theaters, netizen's rating before the release, and numbers of news articles were meaningful. For week 2, the number of screening theaters was relevant for Korean box-office while the ratings and the number of news articles in the first week of release were relevant in the case of foreign box-office. After week 3, , the number of screening theaters, rating and number of news articles from week 2 were relevant for Korean movies, while only the number of screening theaters and number of articles from week 2 were meaningful for foreign movies [3].

Yu Jin, Jungsoo Kim, Jonwoo Kim(2014) analyzed the influence of viral factors among online communities toward box-office that have not been used as a variable in existing research. In order to do so, they concentrated on the subjects such as how size, direction, and network centrality of viral changes as the time after release changes. The analyzed result showed that the network centrality was the most useful factor to measure the influence within the movie communities. Based on this, the viral factors throughout internet communities have been revealed to have gained importance over the other indicators for the film's success in the past [4].

Hoe-Yun Jeong and Hyung-Jeong Yang(2013) analyzed the film success indicators by using multiple regression analysis. Compared to the existing methods of analyzing the film success, their method using multiple regression analysis turned out to be 8.2% more effective at predicting the success of a film. In addition, prediction using artificial neural network turned out to be the most effective method, as it showed89.6% success rate [5].

Kyung Jae Lee and Woo Jin Jang (2006) predicted film success by using Bayesian selection model. In order to do so, they created Bayesian selection model by adding variables reflecting viral effect, release of competing movies along with the differences between the movies and uncertainty of variables, and compared it to existing artificial neural network model. The result appeared to be similar to that of the artificial neural network, but in cases of predicting commercially successful movies, the Bayesian selection turned out to be a superior model. Through these, the most important factors influencing film success turned out to be intensity of competition during the opening week and number of screening theaters in the opening week, *etc.* Actors and actresses, season, movie rating were not very relevant in whole level [6].

Table 1. Previous Research on the Success of Film Prevention

Byoung-Sun Kim(2009)	Analyzed the characteristics of movies based on the way it is released and the screening period, and their influence on the total number of audiences by categorizing the movie into types based on these factors. As a result, it was possible to perceive meaningful difference between the 'Wide release short period' type and 'Narrow release long period' type of movies.
Sun Ju Kwon(2014)	Predict the box office results by analyzing the number of articles from the media and NAVER movie ratings data. As most of the sales occur within the first 3 weeks of opening date, the analysis was done by dividing the period as pre-release, opening week, week 2, and week 3. Analysis showed that relevant variables change dependent on each time period.
Yu Jin, Jungsoo Kim, Jonwoo Kim(2014)	Analyzed the influence of viral factors among online communities toward box-office that have not been used as a variable in existing research. The analyzed result showed that the network centrality was the most useful factor to measure the influence within the movie communities.
Hoe-Yun Jeong and Hyung-Jeong Yang(2013)	Analyzed the film success indicators by using multiple regression analysis. Using multiple regression analysis turned out to be 8.2% more effective at predicting the success of a film.
Kyung Jae Lee and Woo Jin Jang (2006)	Predicted film success by using Bayesian selection model. The result appeared to be similar to that of the artificial neural network, but in cases of predicting commercially successful movies, the Bayesian selection turned out to be a superior model.

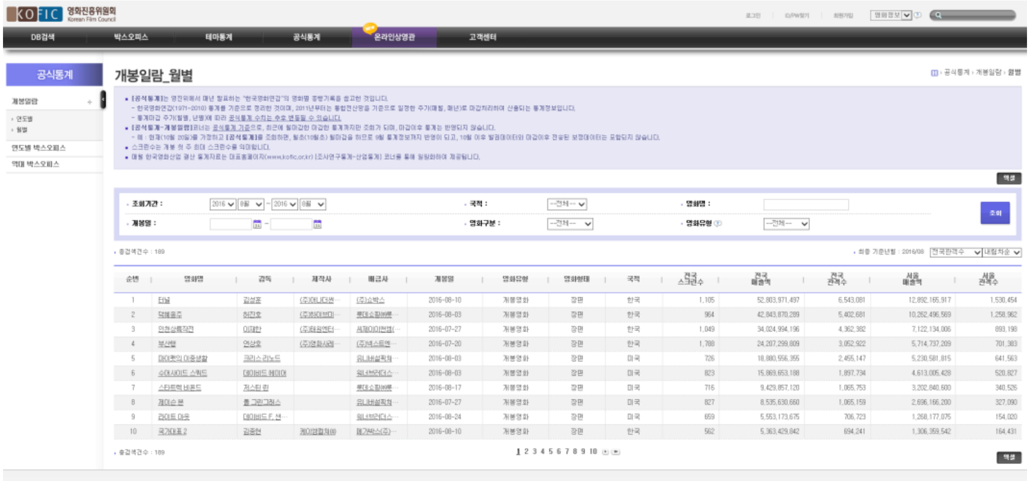
2.2. Random Forest

As a kind of ensemble learning method used in classification, regression analysis, 2.2 Random Forest is a learning mechanism that operates by outputting classification or average predictive value from multiple decision tree composed in training course. Random Forest method is largely composed of learning stage that organizes multiple decision tree and test stage to classify or predict when input vector comes in. Random Forest is being used as various applications such as detection, classification and sessions. As a technique widely used in machine learning, decision making tree can have high discernment as it could deeply grow upon tree characteristic. However, in wrong cases, it has a problem to cause overfitting. Random Forest is a model to reduce errors by leveling such overfitting with creation of several trees. Initial development of Random Forest received influence from an idea to search random subset for the decision available in the context to expand a single tree. The current concept of Random Forest was made of Leo Breiman‘s thesis [8]. This thesis suggested the method to compose forest with trees having no correlation by combining random not optimization and bootstrap aggregating, bagging. A tree in Random Forest is composed of nodes and edges in hierarchical structure. Nodes are divided into internal nodes and longitudinal nodes. Unlike graph, tree is limited that all nodes have only one incoming edge. The number of outgoing edge from each internal node has no limitation, but it is mainly assumed that it has two outgoing edges. As a tree used to make decision literally, decision tree is a technique to divide a complicated question into hierarchical structure type composed with simple questions. Although users can directly set up parameter for a simple question, in case of a complicated question, tree structure and parameter is automatically learned from learning data.

3. Data Analysis

3.1. Data Collection and Settlement

In order to analyze the movie data, the list of box-office ranking data provided by Korean Film Council (http://www.kobis.or.kr) was downloaded. The criteria is nationality and classification of movie, downloaded data from January 2010 to June 2016.



개봉일람_월별

- [공식통계]는 영진위에서 매년 발표하는 "한국영화연감"의 영화별 흥행기록을 참고한 것입니다.
  - 한국영화연감(1971~2010) 통계를 기준으로 정리한 것이며, 2011년부터는 통합전산망을 기준으로 일정한 주기(매월, 매년)로 마감처리하여 산출되는 통계정보입니다.
  - 통계마감 주기(월별, 년별)에 따라 공식통계 수치는 추후 변동될 수 있습니다.
- [공식통계-개봉일람]코너는 공식통계 기준으로, 최근에 월마감한 마감한 통계까지만 조회가 되며, 마감이후 통계는 반영되지 않습니다.
  - 예 : 현재(10월 20일)를 가정하고 [공식통계]를 조회하면, 월초(10월초) 월마감을 하므로 9월 통계정보까지 반영이 되고, 10월 이후 발권데이터와 마감이후 전송된 보정데이터는 포함되지 않습니다.
- 스크린수는 개봉 첫 주 최대 스크린수를 의미합니다.
- 매월 한국영화산업 결산 통계자료는 대표홈페이지(www.kofic.or.kr) [조사연구통계-산업통계] 코너를 통해 일원화하여 제공됩니다.

조회기간 : 2016 8월 ~ 2016 8월
국적 : --전체--
영화명 :
개봉일 : ~
영화구분 : --전체--
영화유형 : --전체--

총검색건수 : 189

최종 기준년월 : 2016/08 전국관객수 내림차순

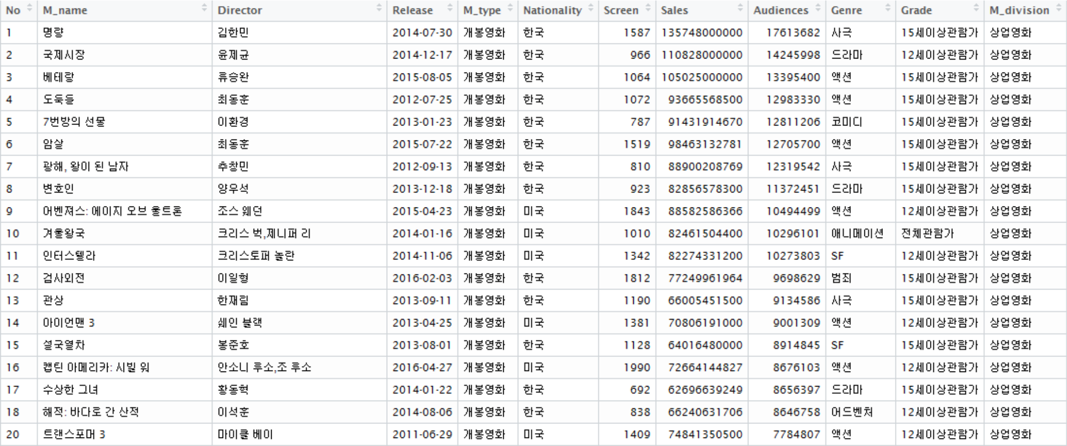
순번	영화명	감독	제작사	배급사	개봉일	영화유형	영화형태	국적	전국 스크린수	전국 매출액	전국 관객수	서울 매출액	서울 관객수
1	터널	김성훈	(주)어나더썬…	(주)쇼박스	2016-08-10	개봉영화	장편	한국	1,105	52,803,971,497	6,543,081	12,892,165,917	1,530,454
2	덕혜옹주	허진호	(주)하이브미…	롯데쇼핑(주)롯…	2016-08-03	개봉영화	장편	한국	964	42,843,870,289	5,402,681	10,262,496,569	1,258,962
3	인천상륙작전	이재한	(주)태원엔터…	씨제이이앤엠(…	2016-07-27	개봉영화	장편	한국	1,049	34,024,994,196	4,362,382	7,122,134,006	893,198
4	부산행	연상호	(주)영화사레…	(주)넥스트엔…	2016-07-20	개봉영화	장편	한국	1,788	24,207,299,809	3,052,922	5,714,737,209	701,383
5	마이펫의 이중생활	크리스 리노드		유니버설픽쳐…	2016-08-03	개봉영화	장편	미국	726	18,880,556,355	2,455,147	5,230,581,815	641,563
6	수어사이드 스쿼드	데이비드 에이어		워너브러더스…	2016-08-03	개봉영화	장편	미국	823	15,869,653,188	1,897,734	4,613,005,428	520,827
7	스타트렉 비욘드	저스틴 린		롯데쇼핑(주)롯…	2016-08-17	개봉영화	장편	미국	716	9,429,857,120	1,065,753	3,202,840,600	340,526
8	제이슨 본	폴 그린그래스		유니버설픽쳐…	2016-07-27	개봉영화	장편	미국	827	8,535,630,660	1,065,159	2,696,166,200	327,090
9	라이트 아웃	데이비드 F. 샌…		워너브러더스…	2016-08-24	개봉영화	장편	미국	659	5,553,173,675	706,723	1,268,177,075	154,020
10	국가대표 2	김종현	케이엠컬쳐(주)	메가박스(주)…	2016-08-10	개봉영화	장편	한국	562	5,363,429,842	694,241	1,306,359,542	164,431

총검색건수 : 189

1 2 3 4 5 6 7 8 9 10

Figure 1. Site of Kobis

The number of total data was 15,472 cases, but the actual number of data used for this analysis was 7,843 cases since the cases with unknown release dates were excluded. In one case, all the information such as Ranking (No), Movie title (M_name), Movie director (Director), Date of release (Release), Movie type (M_type), Nationality (Nationality), Number of screens (Screen), Sales (Sales), Number of audiences (Audiences), Genre(Genre), Grade (Grade), Movie division (M_division) are prescribed.



No	M_name	Director	Release	M_type	Nationality	Screen	Sales	Audiences	Genre	Grade	M_division
1	명량	김한민	2014-07-30	개봉영화	한국	1587	135748000000	17613682	사극	15세이상관람가	상업영화
2	국제시장	윤제균	2014-12-17	개봉영화	한국	966	110828000000	14245998	드라마	12세이상관람가	상업영화
3	베테랑	류승완	2015-08-05	개봉영화	한국	1064	105025000000	13395400	액션	15세이상관람가	상업영화
4	도둑들	최동훈	2012-07-25	개봉영화	한국	1072	93665568500	12983330	액션	15세이상관람가	상업영화
5	7번방의 선물	이환경	2013-01-23	개봉영화	한국	787	91431914670	12811206	코미디	15세이상관람가	상업영화
6	암살	최동훈	2015-07-22	개봉영화	한국	1519	98463132781	12705700	액션	15세이상관람가	상업영화
7	광해, 왕이 된 남자	추창민	2012-09-13	개봉영화	한국	810	88900208769	12319542	사극	15세이상관람가	상업영화
8	변호인	양우석	2013-12-18	개봉영화	한국	923	82856578300	11372451	드라마	15세이상관람가	상업영화
9	어벤져스: 에이지 오브 울트론	조스 웨던	2015-04-23	개봉영화	미국	1843	88582586366	10494499	액션	12세이상관람가	상업영화
10	겨울왕국	크리스 벅,제니퍼 리	2014-01-16	개봉영화	미국	1010	82461504400	10296101	애니메이션	전체관람가	상업영화
11	인터스텔라	크리스토퍼 놀란	2014-11-06	개봉영화	미국	1342	82274331200	10273803	SF	12세이상관람가	상업영화
12	검사외전	이일형	2016-02-03	개봉영화	한국	1812	77249961964	9698629	범죄	15세이상관람가	상업영화
13	관상	한재림	2013-09-11	개봉영화	한국	1190	66005451500	9134586	사극	15세이상관람가	상업영화
14	아이언맨 3	쉐인 블랙	2013-04-25	개봉영화	미국	1381	70806191000	9001309	액션	12세이상관람가	상업영화
15	설국열차	봉준호	2013-08-01	개봉영화	한국	1128	64016480000	8914845	SF	15세이상관람가	상업영화
16	캡틴 아메리카: 시빌 워	안소니 루소,조 루소	2016-04-27	개봉영화	미국	1990	72664144827	8676103	액션	12세이상관람가	상업영화
17	수상한 그녀	황동혁	2014-01-22	개봉영화	한국	692	62696639249	8656397	드라마	15세이상관람가	상업영화
18	해적: 바다로 간 산적	이석훈	2014-08-06	개봉영화	한국	838	66240631706	8646758	어드벤처	12세이상관람가	상업영화
20	트랜스포머 3	마이클 베이	2011-06-29	개봉영화	미국	1409	74841350500	7784807	액션	12세이상관람가	상업영화

Figure 2. Completed Data

3.2. Correlation Analysis

Before predicting the number of audiences, the analysis of correlation between most influential factors for total domestic sales was performed. Total domestic sales were designated as dependent variables, while total number of screening theaters, audiences, and the date of release were designated as independent variables [7].

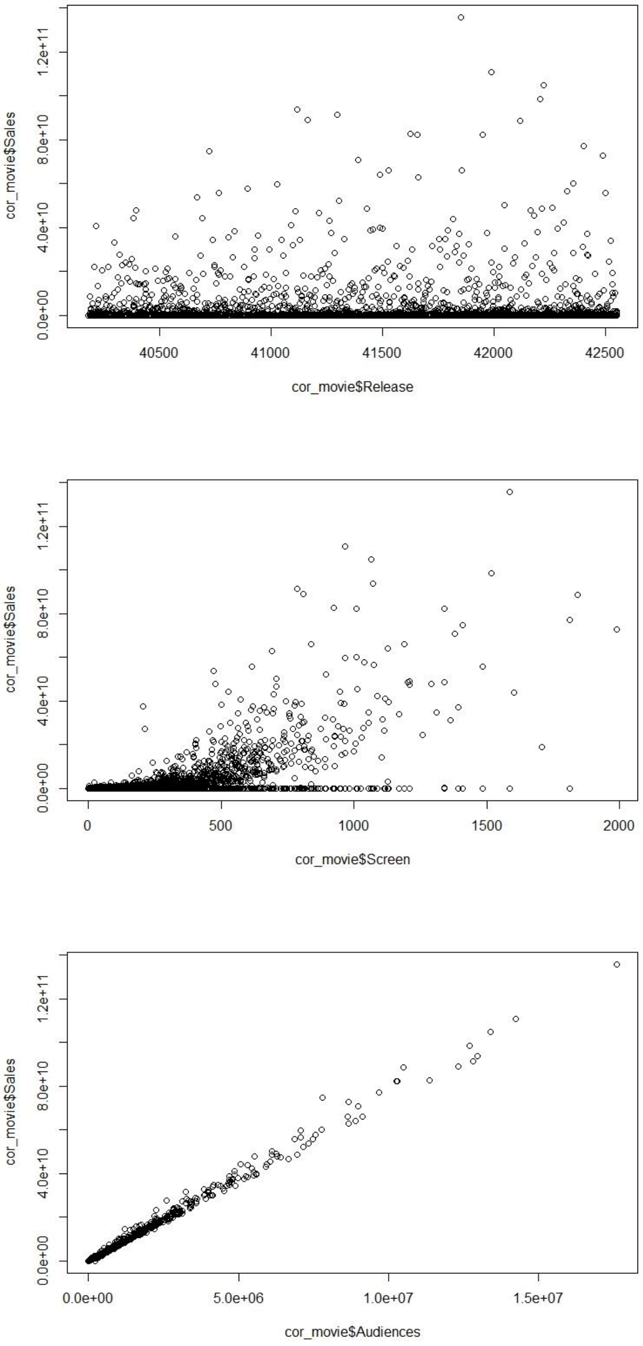


Figure 3. Scattered Chart and Sales Variables

To view the graph in Figure 3 a correlation coefficient is represented as in Table 2.

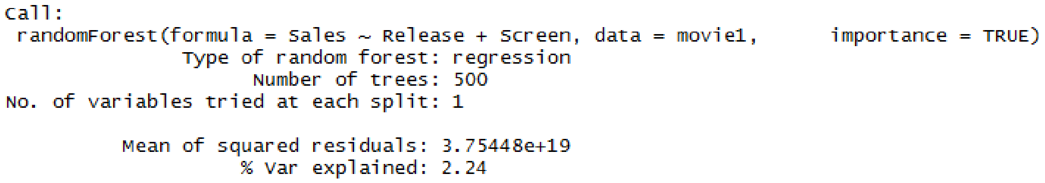
Table 2 Correlation Analysis

	Release	Screen	Sales	Audiences
Release	1.00000000	-	-	-
Screen	-0.02731396	1.00000000	-	-
Sales	-0.04393214	0.58114608	1.00000000	-
Audiences	-0.04662930	0.57688612	0.99772413	1.00000000

Since correlation is only represented in digital data when it comes to correlation analysis, the release date have been changed from date to numbers format. As a result, the variables that were found to be related to the total sales turned out to be screening theaters(0.58) and audiences(0.99). With these values, it has been shown that the number of spectators has the biggest influence on sales of films, with less influence of the number of screens on sales amount. Since it is assumed that the sales increase when the number of audience increases, the number of audience was excluded from 3.3 Random Forest analysis.

3.3. Random Forest Analysis

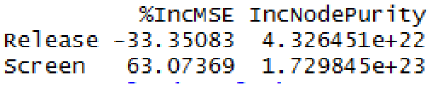
Random Forest algorithm and R program were used for analysis. Random Forest is embodied by several decision-making trees, and randomForest() function is used for measuring the significance of each variables and selecting the variables for modeling. Significance of each variables are measured based on how much each variable contribute to accuracy and Node Impurity improvement.



```
Call:
 randomForest(formula = Sales ~ Release + Screen, data = movie1,      importance = TRUE) 
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 3.75448e+19
                    % Var explained: 2.24
```

The release date and the number of screening theaters were set as independent variables, while sales were set as dependent variable. Figure 1 was used as data for analysis, selected number of trees was set at 500, and the significance of the variables were also analyzed.



```
           %IncMSE IncNodePurity
Release -33.35083  4.326451e+22
Screen   63.07369  1.729845e+23
```

As a result of evaluating each variable’s significance via randomForest by using Importance() function, it ranked each variable’s significance for sales in the order of Number of screening theaters>release date. As for the significance of each variable, Importance type 1 was represented with %IncMSE, and type 2 was represented with Node Impurity (IncNodePurity). %IncMSE is the most robust and informative measure. It is the increase in MSE of prediction as a result of variable being permuted. Graph of significance of each variable using varImpPlot() is shown to be similar to Figure 4.

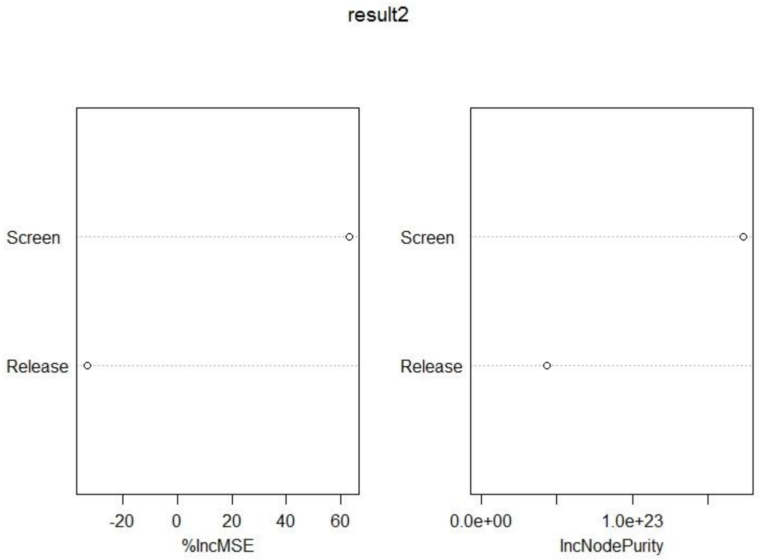


Figure 4. Graph of Significance of Each Variable

4. Conclusion

As predicting the film's success is gaining an increasing importance, many film success prediction studies are performed using various methods. Byeong Sun Kim (2009) categorized movies into two ways; the way it is released and screening period. The characteristics of the movies were determined based on the category, and their effects on the number of audience were analyzed. Sun Ju Kwon(2014) did an analysis based on the number of news articles on the media and NAVER movie rating data. Other studies, have analyzed the viral effect on the internet community, which aren't typically used as a factor in the previous research. Thus, this paper used the data of box-office rankings from January 2010 to Jun, 2016 offered by Korean Film Council in order to predict the success of film. In one case, all the box-office information like Ranking (No), Movie title (M_name), Movie director (Director), Date of release (Release), Movie type (M_type), Nationality (Nationality), Number of screens (Screen), Sales (Sales), Number of audiences (Audiences), Genre(Genre), Grade (Grade), Movie division (M_division) are prescribed. Domestic sales were designated as dependent variable, and its correlation with other variables that might potentially affect sales were analyzed. The result showed that release date(-0.04), number of screening theaters (0.58), and the number of audiences(0.99) were relevant. Through the result of statistical analysis, it has been verified that the influence of capital, which determines the number of screen, is great, and this means that the capital represented by giant distributor is an important factor to be considered in analyzing and understanding the film market. However, as we can see that the correlation between nationwide sales and the number of screen is 0.58, the film having many numbers of screening does not necessarily succeed in box office hit. Through this we can infer that, when a film is not the one over certain level that spectators can universally be satisfied, it is not easy to have a box office hit, though having many number of screening.

The number of audience was excluded among the variables for Random Forest analysis, since it was obvious that the sales increases when number of audiences increases. By using Random Forest algorithm, the relevance of the variables on the total sales were ranked in the order of Number of screens>Release date.

This research was performed using digitized data, and such release data is shown to have negative correlation, as it was converted from date into numeric format. Further research on application of analysis models after pretreatment with Random Forest algorithm in addition to research on predicting success of film through connected models are necessary

References

[1] Industrial policy research team of KOFIC, “2015 Korea film industry settlement”, KOFIC, **(2015)**.

[2] B. S. Kim, “Comparison of Factors Predicting Theatrical Movie Success: Focusing on the Classification by the Release Type and the Length of Run”, Korean Journal of Journalism & Communication Studies, vol. 53, no. 1, **(2009)**, pp.257-287.

[3] S. J. Kwon, “Analysis and Forecasts of movie box office results- Data use of news and web site”, Review of Cultural Economics, vol. 17, no. 1, **(2014)**, pp.35-55.

[4] Y. Jin, J. Kim and J. Kim, “Product Community Analysis Using Opinion Mining and Network Analysis - Movie Performance Prediction Case”, Journal of Intelligent Information Systems, vol. 20, no. 1, **(2014)**, pp. 49-65.

[5] H. Y. Jeong and H. J. Yang, “Predicting Financial Success of a Movie Using Multiple Regression Analysis”, Proceedings of the Korean Society of Computer Information Conference”, vol. 21, no. 2, **(2013)**, pp.275-278.

[6] K. J. Lee and W. J. Jang, “Predicting Financial Success of a Movie Using Bayesian Choice Model”, Industrial Engineering & Management Systems Conference, **(2006)**, pp.1428-1433.

[7] H. K. Lee, H. J. Lee, Y. L. Choi, J. Park, J. Choi and J. B. Kim, “A Study on Correlation Analysis Between CCTV installed Area and CPTED established District”, 2016 International conference on future information & communication engineering, vol. 8, no. 1, **(2016)**, pp.302-303.

[8] L. Breiman, “Random Forest”, Machine Learning, vol. 45, no. 1, **(2001)**, pp. 5-32.

Authors

Hyeon-Kyung Lee, received her bachelor's degree of Computer Information in Baewha Women’s University, Seoul (2015). She is studying her master's degree of software engineering in Graduated Soongsil University, Seoul. Her current research interests include Software engineering and Open source software.

Hong-Jae Lee, received his bachelor's degree of Electronic Engineering in Hanyang University, Seoul (1984). He is studying his Docter's degree in Department of IT Policy and Mgmt., Graduate School of Soongsil University, Seoul. His current research interests include Software engineering and Open source software.

Jeawon Park, received the Ph.D. degree in Computer Science from Soongsil University in Korea, 2011. He is a profressor at Graduate School of Software, Soongsil University. His research interests are in areas of Software Testing, Software Process, Web Services, and Project Management.

Jaehyun Choi, received the Ph.D. degree in Computer Science from Soongsil University in Korea, 2011. He is a profressor at Graduate School of Software, Soongsil University. His research interests are in areas of Data Processing, Service Engineering, Software Engineering, and Text Mining.

Jong-Bae Kim, received his bachelor's degree of Business Administration in University of Seoul, Seoul (1995) and master's degree (2002), doctor’s degree of Computer Science in Soongsil University, Seoul (2006). Now he is a professor in the Graduate School of Software, Soongsil University, Seoul, Korea. His research interests focus on Software Engineering, and Open Source Software.