International Journal of Forecasting 31 (2015) 364–390

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Box office forecasting using machine learning algorithms based on SNS data

Taegu Kim [a], Jungsik Hong [b], Pilsung Kang [c,*]

[a] Industrial Engineering, Seoul National University, Republic of Korea

[b] Industrial & Information Systems Engineering, Seoul National University of Science and Technology, Republic of Korea

[c] School of Industrial Management Engineering, Korea University, Republic of Korea

ARTICLE INFO

Keywords:
Box office earning forecast
Social network service
Machine learning
Genetic algorithm
Forecast combination

ABSTRACT

We propose a novel approach to the box office forecasting of motion pictures using social network service (SNS) data and machine learning-based algorithms. We begin by providing a comprehensive survey of the forecasting algorithms and explanatory variables used in the motion picture domain. Because of the importance of forecasting in early periods, we develop three sequential forecasting models for predicting the non-cumulative and cumulative box office earnings: (1) prior to, (2) a week after, and (3) two weeks after release. The numbers of SNS mentions and their weekly trends are used as input variables in addition to the screening-related information. A genetic algorithm is adopted for determining significant input variables, whereas three machine learning-based nonlinear regression algorithms and their combinations are employed for building forecasting models. Experimental results show that the utilization of SNS data, machine learning-based algorithms and their combination made noticeable improvements to the forecasting accuracies of all the three models.

1. Introduction

Globally, the motion picture industry is one of the most rapidly growing industries. Thousands of new movies are released every year, and the compounding annual growth rate of the film market is expected to be as high as 5.88% between 2010 and 2015.[1] Because the public's taste is unpredictable, newly released motion pictures are usually exposed to a high risk of meeting the break-even point. For this reason, box office forecasting has always been a major concern in the motion picture business (Jun, Kim, & Kim, 2011). Needless to say, having an accurate box office forecast in an early stage of screening is extremely important, because marketing activities in this period decide the final success of motion pictures (Sawhney & Eliashberg, 1996). In addition, producers and exhibitors can benefit from accurate forecasts, which would allow them to make appropriate managerial decisions concerning budget allocations for additional marketing activities and screen allocation across movie theaters.

To date, a number of studies have attempted to develop effective pre-release or early stage box office forecasting models; these studies can be divided into two categories based on their focus: (i) forecasting algorithm development or (ii) explanatory variable configuration. For those studies which focus on algorithm development, the following three subgroups exist: (1) statistical learning models, including linear regression and probabilistic models (Chintagunta, Gopinath, & Venkataraman, 2010; Eliashberg, Jonker, Sawhney, & Wierenga, 2000; Eliashberg & Shugan, 1997; Litman, 1983; Ravid, 1999; Sawhney & Eliashberg, 1996); (2) time series forecasting models such as new

* Corresponding author.
E-mail addresses: pilsung.kang@gmail.com, pskang@seoultech.ac.kr (P. Kang).

[1] PWC, Global Entertainment and Media Outlook: 2011–2015.

http://dx.doi.org/10.1016/j.ijforecast.2014.05.006
0169-2070/

product diffusion models and the vector auto-regression (VAR) method (Dellarocas, Zhang, & Awad, 2007; Lee, Kim, & Cha, 2012; Rogers, 1976; Wang, Zhang, Li, & Zhu, 2010); and (3) machine learning-based models such as artificial neural networks (ANN) (Delen, Sharda, & Kumar, 2007; Zhang, Luo, & Yang, 2009). Meanwhile, for those studies which focus on the explanatory variable configuration, the main topic is the way in which movie characteristics and word-of-mouth (WOM) can be incorporated into explanatory variables. Star/director/distributor power, genre, and degree of competition are commonly exploited as movie characteristic-based input variables (Delen et al., 2007; Litman, 1983; Ravid, 1999; Wen & Yang, 2011), whereas the volume and level of user ratings, or the volume and sentimental polarity of blog posts or SNS mentions, are tested as WOM-related input variables (Asur & Huberman, 2010; Dellarocas et al., 2007; Duan, Gub, & Whinston, 2008; Liu, 2006; Qin, 2011).

Although box office forecasting in the motion picture industry has been addressed widely, some limitations still exist. The first issue is the deficiency of forecasting algorithm diversity. So far, statistical regression algorithms such as multivariate linear regression (MLR) have prevailed in the research on box office forecasting because they can explain the influence of each variable on the forecast (Chintagunta et al., 2010; Elberse & Eliashberg, 2003; Eliashberg & Shugan, 1997; Litman, 1983; Ravid, 1999; Simonoff & Sparraw, 2000). Because this type of algorithm assumes a linear relationship between the input and target variables, its prediction accuracy is usually inferior to that of algorithms that can capture both linear and nonlinear relationships. Time series based algorithms such as the Bass diffusion models can relieve the linearity assumption (Dellarocas et al., 2007; Lee et al., 2012; Wang, Zhang, et al., 2010), but since they use only the historical sales records, the forecast improvement is limited because the demand for motion pictures is also affected by various exogenous variables. Although a few studies have intended to develop forecasting models based on nonlinear algorithms such as ANN, there is a considerable scope for performance improvement if various machine learning-based regression algorithms are taken into consideration.

The second issue is associated with variable configuration. Although there is an ongoing claim that movie characteristics play a central role in box office forecasting, the variable configurations usually rely on the subjective judgments of human experts, so that there is no universally accepted specification for any of them (Ainslie, Drèze, & Zufryden, 2005; Elberse & Eliashberg, 2003; Lee & Chang, 2009; Litman, 1983; Liu, 2006; Qin, 2011; Ravid, 1999; Wen & Yang, 2011). For example, star power is sometimes designed as a binary variable that is assigned the value 1 if the main actor satisfies a list of certain conditions and is 0 otherwise (Chintagunta et al., 2010; Litman, 1983). In other cases, star power is designed as an ordinal or continuous variable that reflects the number of audiences or the amount of box office earnings of the previous movie in which he/she appeared (Ainslie et al., 2005; Delen et al., 2007; Lee & Chang, 2009). Moreover, contradictory results are reported occasionally; star power was found to be significant in certain cases (Chintagunta et al., 2010; Delen et al., 2007), but insignificant in other cases (Ainslie et al., 2005; Lee & Chang, 2009; Litman, 1983). In summary, the significance of a movie characteristic in a forecasting model depends on not only what the characteristic itself represents, but also the way in which modelers specify it, which is too case-specific, meaning that the same configuration cannot be applied to other cases directly.

There are two main directions possible for resolving the aforementioned issues: increasing either the data diversity or the model complexity. We are currently living in the era of big data sets, where the amount of data is increasing explosively, and virtually unlimited computational power is available for processing these large data sets. The box office records, together with other screening-related data from the entire market, are updated and managed by credible agencies on a daily basis in many countries, e.g., Korea Film Council (KOFIC) in Korea, while various types of personal preference data on motion pictures are available on the web. User ratings can be collected from motion picture rating websites such as IMDB[2] in the United States or NAVER[3] in Korea. User comments, whether informative or emotional, can be collected from internet communities, personal blogs, and social network services (SNSs). As has been noted, WOM plays a crucial role in forecasting motion picture successes. Until now, WOM data have usually been obtained by conducting customer interviews, random dialing surveys, or on-line user ratings, or even by conducting a simulation (Dellarocas et al., 2007; Duan et al., 2008; Eliashberg et al., 2000; Mestyán, Yasseri, & Kertész, 2013; Zufryden, 1996). Not only do these methods require a lot of effort at a high cost, they also suffer from sampling bias, which would not reflect public interests properly. With the emerging prevalence of SNS, it has become possible to collect transparent, authentic, and consumer-driven WOM data on motion pictures (Abel, Diaz-Aviles, Henze, Krause, & Siehndel, 2010; Asur & Huberman, 2010; Mestyán et al., 2013; Qin, 2011). Furthermore, a number of websites provide refined and analyzed SNS data, such as Google Blog Search[4] and BlogPulse,[5] which mitigate the extensive efforts of text preprocessing. By incorporating the WOM data collected from SNSs, we expect that more influential explanatory variables will be identified, relative to conventional movie descriptive variables.

Another improvement can be achieved by adopting sophisticated machine learning based nonlinear regression algorithms. Because most forecasting models are based on linear algorithms, they cannot determine nonlinear relationships that may increase the forecasting accuracy. As has been mentioned, only a few trials have attempted to develop forecasting models on the basis of nonlinear algorithms; there has been no such work on exploiting diverse machine learning algorithms for forecasting model development in the motion picture domain. The employment of various well-designed nonlinear regression algorithms, such as support vector regression (SVR), Gaussian process regression (GPR), and k-nearest neighbors (k-NN), and the

[2] http://www.imdb.com.
[3] http://movie.naver.com/.
[4] http://blogsearch.google.com.
[5] http://blogpulse.com.

combination of individual forecasting models, can enhance the forecasting accuracy.

In this paper, we aim to develop three serial box office forecasting models in the early stages of screening, with the help of SNS-related data and sophisticated machine learning-based regression algorithms. The first forecast is made immediately before release, and two subsequent forecasts are made one and two weeks after release. Both screening-related information and SNS-derived data are considered as input variables; the former includes the number of seats and accumulated box office sales, whereas the latter includes the total number of informative/emotional/positive/negative mentions on a weekly basis and their weekly/overall trends collected from various SNS services. The forecasting accuracy of the conventional model is first enhanced by employing machine learning based regression algorithms, then boosted further by combining the forecasts of individual models. We believe that our forecasting models are not only accurate and easily reproducible from a modeling perspective, but also of practical application, because neither human experts-oriented variable configurations nor extensive data preprocessing is required.

The remainder of this paper is structured as follows. Section 2 provides a comprehensive survey of the literatures associated with box office forecasting in the motion picture industry, with a particular focus on forecasting algorithms and explanatory variables. Section 3 demonstrates the overall research framework, including data collection, variable selection, use of a single/combined forecasting model, and the performance criterion. Section 4 validates our forecasting models by decomposing the overall improvement into the usage of SNS data, the adoption of machine learning algorithms, and the combination of individual models' forecasts. In Section 5, we discuss limitations of the current work and suggest future research directions, as well as providing concluding remarks.

2. Literature review

There are two main streams in the literature on the field of box office forecasting models, which focus on: (i) developing accurate forecasting algorithms and (ii) exploiting appropriate explanatory variables. Here, we briefly review the representative studies in each category and discuss their significance and limitations.

2.1. Forecasting algorithms

From a forecasting algorithm-centric viewpoint, the following four major categories can be found, as summarized in Table 1: (1) statistical learning based models, (2) probabilistic models, (3) diffusion based models, and (4) machine learning based models. Linear regression was the algorithm employed most commonly among statistical learning based models. In Litman's (1983) work, which is one of the pioneering studies that provided grounds for the econometric analysis of motion pictures, movie characteristics such as the type of story, Motion Picture Association of America (MPAA) ratings, star, cost, and major distributor were considered as input variables, whereas the theatrical rentals accruing to the distributor were set as the target variable. Using linear regression as a base model, the forecasts made prior to release reported an R^2 value of 0.485. Eliashberg and Shugan (1997) and Ravid (1999) also used similar approaches, with the only difference being the input variable selection. The former used the number of screens and total/positive/negative critics' reviews, whereas the latter used the star power, budget, sequels, and MPAA ratings. Both of these studies evaluated the forecasting models in terms of R^2, and reported slightly better performances than Litman (1983). Chintagunta et al. (2010) built five market-level or national-level forecasting models by adopting linear regression on the basis of WOM information and movie characteristics. They reported R^2 values of between 0.9344 and 0.9892. These performances may appear remarkable, but their models were only intended to forecast the opening day box office, not the total box office, which restricts the utility of the forecasts.

Probabilistic models were introduced as an alternative to linear regression-based models. Sawhney and Eliashberg (1996) developed a forecasting model named BOXMOD-I, which is based on queuing theory; it assumes three box office patterns, namely the exponential distribution, Erlang-2 distribution, and generalized Gamma distribution, and forecasts are updated weekly for the first three weeks by adapting the information collected during each week. Although the forecasting error on the day of release was very high, 71% in terms of mean absolute percentage errors (MAPE), it decreased significantly to 7.2% when the cumulative box office data over the first three weeks is included. However, it cannot be accepted on a general basis because only 19 motion pictures were analyzed. Eliashberg et al. (2000) developed MOVIEMOD, which is an enhanced version of BOXMOD-I, by taking into account behavioral variables such as the WOM frequency, WOM duration, and distribution delay, as well as movie characteristics. Although the suggested framework is logically sound, its practicality was not verified sufficiently, since only one motion picture was analyzed and the WOM data were generated artificially by simulation.

Diffusion-based models have prevailed among the time series based forecasting models. The goal of the diffusion model is to explain the spread pattern of a new product or service in a social system by determining how it is perceived by the customers (Rogers, 1976). Because a considerable proportion of the total box office earnings are generated soon after release, there was an attempt to apply an exponential decay model to box office forecasting. Jedidi, Krider, and Weinberg (1998) divided 102 movies into four clusters based on the estimates of the parameters in the exponential decay model. The resulting clusters were found to differ from each other with respect to various attributes such as genre, ratings, stars, awards, seasonality, and competition. The Bass diffusion model is another popular diffusion-based forecasting method. Dellarocas et al. (2007) produced box office forecasts by employing a variation of the Bass diffusion model. Movie characteristics such as genre, MPAA ratings, and star power, together with user review data collected from two major movie websites,

Table 1
The forecasting algorithms and performance metrics used in previous studies.

Category	Research	Algorithm	Target	Performance metric
Statistical model	Litman (1983)	MLR	N/A	R^2
	Zufryden (1996)	MLR	Logged	R^2
	Eliashberg and Shugan (1997)	MLR	N/A	Adjusted R^2
	Vany and Walls (1999)	MLR	Logged	R^2
	Ravid (1999)	MLR	N/A	Adjusted R^2
	Simonoff and Sparraw (2000)	MLR	Logged	R^2, RMSE
	Elberse and Eliashberg (2003)	MLR	Logged	R^2
	Liu (2006)	MLR	Logged	Adjusted R^2
	Duan et al. (2008)	3SLS MLR	Logged	R^2
	Brewer et al. (2009)	MLR	N/A	Adjusted R^2
	Zhang and Skiena (2009)	MLR	Logged	AMAPE
	Abel et al. (2010)	MLR	N/A	MAPE, MAE, correct ratio
	Asur and Huberman (2010)	MLR	N/A	R^2, AMAPE
	Calantone et al. (2010)	MLR (3SLS)	Logged	R^2
	Chintagunta et al. (2010)	MLR	Logged	Adjusted R^2
	Gong et al. (2011)	MLR	N/A	Adjusted R^2
	Qin (2011)	MLR (2SLS)	Logged	R^2
	Wen and Yang (2011)	MLR	N/A	Adjusted R^2
	Lovallo et al. (2012)	MLR	Logged	Percentage relative error
	Marshall et al. (2013)	MLR	Logged	MAPE
	Mestyán et al. (2013)	MLR	Logged	R^2
Probabilistic model	Eliashberg and Sawhney (1994)	ENJMOD	N/A	R^2
	Sawhney and Eliashberg (1996)	BOXMOD	N/A	MSE, MAPE
	Neelamegham and Chintagunta (1999)	Hierarchical Bayes model	N/A	RMSE, MAE
	Eliashberg et al. (2000)	MOVIEMOD	N/A	Percentage relative error
	Ainslie et al. (2005)	Logit model with Gamma distribution	N/A	MAPE
	Jun et al. (2011)	DYNAMIC	N/A	MAD, RMSE
	Marshall et al. (2013)	BOXMOD	N/A	MAPE
Diffusion model	Jedidi et al. (1998)	Exponential decay	N/A	N/A
	Dellarocas et al. (2007)	Bass	N/A	MAPE
	Wang, Zhang, et al. (2010)	Bass	N/A	MAPE
	Lee et al. (2012)	Generalized Bass model	N/A	Adjusted R^2, MAPE
	Marshall et al. (2013)	Bass	N/A	MAPE
Machine learning	Delen et al. (2007)	ANN, DT	N/A	N/A
	Lee and Chang (2009)	BBN, ANN, DT	N/A	Correction ratio
	Zhang et al. (2009)	ANN	N/A	Average percent hit rate
	Abel et al. (2010)	Bayes Net, SMO	N/A	MAPE, MAE, correct ratio

Note: The column titled "Target" denotes the way in which the target variable, mostly box office earnings, is transformed.

were subsequently analyzed after the forecast had been made based on only historical box office records. When the early box office data were not available, their model reported a MAPE of 24%, but this value decreased to 10% when the first three days of box office data became available. Lee et al. (2012) built a forecasting model on the basis of a generalized Bass model, which considered the various seasonal factors and herding effects, as well as external and internal influences. Experimental results showed that the proposed method could achieve a 12.7% average MAPE for 40 motion pictures in the Korean market when two weeks of box office records were available. Marshall, Dockendorff, and Ibáñez (2013) also applied the Bass model to the Chilean film market, but the forecasting performance was not meaningful; no statistically significant differences were found between the Bass model and the linear regression based benchmark models. Although some of the aforementioned studies reported fairly good forecasting performances, the inherent limitation of time series based forecasting models is that their forecasts depend solely on the historical box office records; critical exogenous variables are not taken into consideration.

There has been little progress in the development of machine learning based forecasting models in the motion picture domain. In addition, most of them formulated box office forecasting as a classification problem that predicts whether a specific motion picture will earn more than a certain amount of money, rather than as a regression problem that predicts actual box office earnings. A large amount of information might be lost when the forecasting is formulated as a classification, and this information loss restricts the usability of forecasting results. Delen et al. (2007) developed motion picture success forecasting models by employing ANN, decision trees, and discriminant analysis to classify the box office size. Zhang et al. (2009) also adopted ANN as a base model for classifying the size of the box office into one of six predefined categories. Lee and Chang (2009) attempted to apply a Bayesian belief network to the forecasting of movie success. The purpose of their model was to classify a new movie into one of three classes, depending on the audience numbers. Abel et al. (2010) tried to forecast actual sales, rather than predicting arbitrary manipulated categories, by applying eight different machine learning algorithms based on blog data. These algorithms were compared with the simple linear regression

model for forecasting (1) box office earnings of movies and (2) sales of music albums. For both target variables, the machine learning algorithms performed better than the linear model. In particular, the forecasting accuracies for movies were fairly good, recording a MAPE of 16.41%, which was much lower than that for sales of music albums (26.21%). The main problem with these studies is that, with only a few exceptions, not only are few sophisticated machine learning algorithms considered, but also the target is manipulated arbitrarily, in that the continuous box office is transformed into a few ordinal categories. Once the box office has been manipulated as such, information loss is inevitable, so exhibitors will be unable to make precise managerial decisions.

2.2. Explanatory variables

From a variable-centric viewpoint, most studies focus on the configuration of movie-specific variables. Table A.1 in Appendix A summarizes the ways in which the explanatory variables introduced in previous studies were measured and whether they were found to be significant in forecasting box office earnings. Movie characteristics and WOM are the two main categories that can distinguish these input variables properly. Screen, age of movie, star power, advertising effect, award, budget, critic, director power, distributor power, genre, nationality, rating, sequel, timing of release, and competition are all considered as movie characteristics related variables, whereas awareness and preference are considered as WOM related variables.

Screen is the variable considered most commonly in box office forecasting. Not only does it reflect the supply capacity for movies accurately, but also information regarding this variable can be acquired easily from reliable sources. As we expected, there is a consistent confirmation that screen-related variables are almost always significant, irrespective of the way in which they are measured and the forecasting algorithms used. This variable is sometimes measured by the number of screens on a daily (Qin, 2011) or weekly basis (Calantone, Yeniyurt, Townsend, & Schmidt, 2010; Elberse & Eliashberg, 2003; Liu, 2006; Neelamegham & Chintagunta, 1999), whereas other studies have adopted the number of theaters, not screens, as a measurement unit (Asur & Huberman, 2010; Zufryden, 1996).

The running period after a movie's release is another variable that can be defined objectively. Moreover, information regarding this variable is easy to collect from reliable sources. It is measured on either a daily (Chintagunta et al., 2010; Qin, 2011) or a weekly (Neelamegham & Chintagunta, 1999; Wang, Zhang, et al., 2010; Zufryden, 1996) basis. Since most previous studies have focused on a specific time, such as pre-release or one week after launch, this variable is not typically considered as screen-related information. However, every previous study that considered the running period concluded that the age of the movie played an important role in box office forecasting.

Among movie-related characteristics, genre and rating have been considered commonly as explanatory variables in box office forecasting, because they are the main determinants of the potential market size. Both variables have appeared in many studies in a similar form, namely as dummy variables for each genre and rating group, respectively. Interestingly, Gong, Young, and der Stede (2011) adopted only one genre-derived variable to indicate a hi-tech genre, while Sawhney and Eliashberg (1996) measured the MPAA rating using integer values from one to four. It is hard to come to any general conclusion on the significance of genre and rating, as they have been supported positively by some studies (Brewer, Kelley, & Jozefowicz, 2009; Calantone et al., 2010; Dellarocas et al., 2007; Gong et al., 2011; Simonoff & Sparraw, 2000) and have been found questionable by others (Lee, 2009; Litman, 1983; Lovallo, Clarke, & Camerer, 2012; Qin, 2011).

The star power of actors/actresses is another very widely explored variable. Since it is a fundamentally qualitative attribute, no clear agreement exists as to its quantification. Sawhney and Eliashberg (1996) created a dummy variable to indicate whether an actor/actress is famous enough to possess "marquee value", and this was later adopted by Neelamegham and Chintagunta (1999). In other studies, it has been measured based on whether or not the main role casts were listed in famous magazines (Brewer et al., 2009; Elberse & Eliashberg, 2003; Simonoff & Sparraw, 2000), or the star's historical box office earnings (Gong et al., 2011; Litman, 1983; Lovallo et al., 2012; Ravid, 1999; Simonoff & Sparraw, 2000). Another interesting adoption of the star power in forecast modeling is the creation of a dummy variable that indicates whether the main cast wins at, or is nominated by, authorized film festivals. Because there is neither an academic nor a commercial definition for stars, some researchers have even developed their own criteria (Jun et al., 2011). Due mainly to the differences in definitions, as well as to time and market differences, controversial results have been reported on the significance of star power: of the 11 studies reviewed, three (Elberse & Eliashberg, 2003; Gong et al., 2011; Simonoff & Sparraw, 2000) argued its significance, while the others were against it (Brewer et al., 2009; Jun et al., 2011; Litman, 1983; Lovallo et al., 2012; Neelamegham & Chintagunta, 1999; Ravid, 1999; Sawhney & Eliashberg, 1996; Wen & Yang, 2011).

In addition to star power, the powers of the director and distributor are two other brand-oriented variables that may affect audience expectations. The power of the director has been measured in similar ways to that of the stars, by referring to relevant lists (Elberse & Eliashberg, 2003) or using historical box office earnings (Gong et al., 2011). In some other works, such as that of Wen and Yang (2011), professional critics have been used to evaluate directing power. On the other hand, a very simple and common indicator has also been used to reflect distribution power. It designates whether or not the target movie is distributed by one of the major distribution companies, which are judged subjectively by domain experts, in the target market (Calantone et al., 2010; Gong et al., 2011; Jun et al., 2011; Litman, 1983). A unique and well quantified measure of distribution power was introduced by Wen and Yang (2011), who counted the number of film titles previously

released by the distributor. Despite the high public awareness, previous studies have reported that neither the director nor the distribution power were meaningful predictors of box office earnings for all cases except one (Gong et al., 2011).

Budget and advertising are expenditure-oriented predictors in box office forecasting. The production cost is the only parameter that reflects a movie's budget, while advertising can be measured in more diverse ways, such as advertising frequency (Chintagunta et al., 2010; Wang, Zhang, et al., 2010), number of theaters in the opening week (Lovallo et al., 2012), and the proportion of respondents who intend to watch the movie after viewing its trailer (Eliashberg et al., 2000). Although both budget and advertising can be meaningful predictors in box office forecasting (provided that the information is available), from a practical point of view it is very difficult to collect the relevant information for hundreds of movies from reliable sources.

Critics and awards are reputation-related variables, the information for which is usually available only after press previews, or post release in other countries. In most studies, the reviews appeared in magazines (Elberse & Eliashberg, 2003; Eliashberg & Shugan, 1997), newspapers (Litman, 1983; Simonoff & Sparraw, 2000), and online rating sites (Brewer et al., 2009), which were scraped to create critic-related variables. Eliashberg and Shugan (1997) and Ravid (1999) classified reviews depending on the position taken by the reviewer; the former divided the reviews into three categories (pro, con, and mixed), whereas the latter simply categorized them as either good or bad. On the other hand, Brewer et al. (2009) and Simonoff and Sparraw (2000) adopted raw information from specific sources, i.e., the Rotten Tomato ratings (percentage) and Roger Ebert's ratings, respectively. Of the ten studies reviewed, half argued that professional reviews have a critical effect on public preferences (Brewer et al., 2009; Dellarocas et al., 2007; Elberse & Eliashberg, 2003; Litman, 1983; Liu, 2006), whereas three studies disproved the argument (Ravid, 1999; Simonoff & Sparraw, 2000; Wen & Yang, 2011). The variable award has usually been reflected by creating a dummy variable that indicates whether or not the target movie won an award or was nominated in specific categories (Litman, 1983; Simonoff & Sparraw, 2000; Wen & Yang, 2011). Unlike the two earlier studies (Litman, 1983; Simonoff & Sparraw, 2000), Wen and Yang (2011) insisted on the insignificance of the award effect in the Chinese film market.

Distribution strategy is another factor which has been considered intensively in box office forecasting. It is reflected by either the release timing or the competition environment. First, the timing of release was defined mostly as whether or not the target movie had been released during certain periods, such as a weekend (Duan et al., 2008), summer season (Brewer et al., 2009; Gong et al., 2011; Litman, 1983; Simonoff & Sparraw, 2000), and the Christmas holiday season (Brewer et al., 2009; Gong et al., 2011; Litman, 1983; Simonoff & Sparraw, 2000; Zhang & Skiena, 2009). Two alternative approaches include considering (1) the score on a 1–100 scale of the AC Nielsen EDI (Elberse & Eliashberg, 2003) and (2) a continuous measure for seasonality (Ravid, 1999). Eight of the 13 studies reviewed supported the inclusion of seasonality in the forecasting model, as it improved the predictive power, while the others did not (see Table A.1). However, so far, there is no commonly accepted definition for the competition environment. Gong et al. (2011) simply adopted the rank of the target movie on the opening weekend and converted it to a variable on a scale of 0–3, while Wen and Yang (2011) created an indicator for the existence of a power substitute. More complicated measurements for competition intensity include the average age of newly released films (Liu, 2006), the total number of competing movies in a certain theatrical period (Calantone et al., 2010; Elberse & Eliashberg, 2003; Liu, 2006), and the use of critic scores and the star power of competing movies (Chintagunta et al., 2010). Despite the various attempts to measure the distribution strategy and competition environments, their significances were not recognized as expected; only two studies indicated a positive conclusion (Calantone et al., 2010; Elberse & Eliashberg, 2003).

Various other movie-specific factors, such as income, price, sequel, and nationality, may also affect box office earnings, as is shown in Table A.1. However, the experimental results in the literature are not sufficient to help us to reach any decisive conclusion or recognize any apparent trend in the significance of these factors. Based on our intensive literature survey on the influence of movie-specific factors in box office forecasting, only two variables, i.e., screen and age, have been proven repeatedly to be significant, whereas no such consistent results exist for the other variables. This situation is caused mainly by the absence of standard quantification methods for qualitative information and the differing circumstances of the target film markets and the time periods being analyzed.

WOM-related variables can be divided into two categories: awareness and preference. Awareness is associated with the popularity of a movie with the public, while preference is associated with people's likes and dislikes. In early studies, the degree of recognition among potential audiences was measured indirectly by cumulative viewership (Neelamegham & Chintagunta, 1999) or even through a simulated experiment (Eliashberg et al., 2000). Later, the emerging internet technology provided researchers with access to more direct sources of WOM; early attempts utilized user activities, such as the numbers of user reviews (Chintagunta et al., 2010; Duan et al., 2008), ratings (Dellarocas et al., 2007), or total posts (Wang, Cai, & Huang, 2010), from centralized web sites. Dellarocas et al. (2007) used user review data collected from two movie portal sites, namely Yahoo! Movies[6] and BoxOfficeMojo,[7] to reflect the public awareness in forecasting. In that study, the correlation between user ratings and critics was very low, but using both of them did improve the forecasting accuracy. In addition, the volumes of weekly user ratings and weekly box office earnings were found to be highly correlated. Similar results were also reported by Duan et al. (2008) and Qin (2011). Duan et al. (2008) analyzed the relationship between user review data and the box office earnings and concluded that it is the total volume of ratings

[6] www.movies.yahoo.com.

[7] www.boxofficemojo.com.

that leads to a large box office, not high average ratings. Qin (2011) used more primitive WOM data than user ratings, i.e., the exposure frequency of a motion picture in personal blogs on the internet, and concluded that the volume and the box office affect each other alternately; a large volume of posts increases the box office in this week, which triggers more posts in the next week. It appears that as SNS data grow exponentially, more authentic and genuine user-driven data become available. Liu (2006) and Qin (2011) used the number of blog messages as an awareness indicator, whereas Asur and Huberman (2010) and Wang, Zhang, et al. (2010) tracked the numbers of SNS posts, such as Twitter mentions, as a measurement of awareness levels. Most of the studies reviewed agreed that public awareness is one of the key determinants in box office forecasting. Moreover, the significance of awareness has been strengthened by recent studies that measured the degree of awareness using the frequency of SNS mentions (Asur & Huberman, 2010; Wang, Zhang, et al., 2010).

In contrast to awareness, incorporating preferences in box office forecasting has a relatively short history. Barring the simulation by Eliashberg et al. (2000), all of the studies we reviewed were conducted after the introduction of the internet, which allowed them direct access to massive amounts of user-created data, such as short reviews, movie portal site ratings, or mentions in popular SNS sites (Asur & Huberman, 2010; Chintagunta et al., 2010; Duan et al., 2008; Liu, 2006; Wen & Yang, 2011). One straightforward measure of public preference is to use the average ratings provided by actual movie viewers (Chakravartya, Liub, & Mazumdarc, 2010; Duan et al., 2008; Wen & Yang, 2011). Since movie rating itself is a quantitative concept, it is usually measured on a 5-star or 10-digit (5 or 7, in some cases) scale; the more stars or the higher the values, the more people like the movie. Thus, the only way to eliminate a potential sampling bias is to collect sufficient numbers of ratings for each movie. For example, Duan et al. (2008) collected an average of 1350 user ratings each for 71 movies. An interesting recent attempt extracted the level of preference from raw texts rather than from explicit ratings (Asur & Huberman, 2010; Liu, 2006). It was based on the assumption that, although some people may participate actively on movie portal sites by posting their reviews and ratings for movies they watched, the majority of 'ordinary' people will not do so. Rather, they may express their thoughts and feelings occasionally by putting up a post on their personal blogs or by leaving a short mention on SNS sites. Consequently, provided that it is possible, analyzing the emotional polarity of written texts can provide a better measure of public preferences, since the numbers of such texts significantly outnumber explicit ratings. Liu (2006) collected 12,136 posts from the Yahoo! movie message board and recruited three human experts to read and classify each post into one of the five following categories: positive, negative, mixed, neutral, and irrelevant. The experimental results demonstrated that positive and negative posts are effective only when forecasting the box office in the first week. Since the classification task in Liu's (2006) study is labor intensive and time consuming, Asur and Huberman (2010) designed a more automated data collection and sentiment analysis framework. They first extracted 2.89 million tweets referring to 24 movies from Twitter.com. Then, they adopted a professional linguistic analysis tool (LingPipe[8]) to conduct sentiment analysis. Their forecasting model, which is based on a simple linear regression, reported an adjusted R^2 value of 0.94 when the frequency of tweets (awareness) and the positive-to-negative ratio (preference) were considered simultaneously.

In summary, the significance of awareness in box office forecasting has been proved consistently, because of a relatively simple and more standardized data collection process. Although preferences genuinely contain more information than awareness, an immature sentiment analysis technology was the main obstacle preventing researchers from including preferences in the forecasting model. Thanks to recent advances in natural language processing technology, however, it is possible to extract accurate preferences from massive amounts of plain text. As was confirmed by Asur and Huberman's (2010) pioneering work, the box office forecasting performance could be enhanced further if preference and awareness are considered together.

3. Methodology

The overall research framework is illustrated in Fig. 1. In step 1, we select motion pictures for further analysis by applying some filtering rules such as the generality of the title words and the total number of audiences. For the selected motion pictures, screening-related and SNS-related data are collected between three weeks prior to release and three weeks after release. In step 2, the structures of three forecasting models are determined: when and what to forecast, and which input variables to use. Two different types of target variables are defined: the weekly box office earnings (Target 1) and the cumulative box office earnings (Target 2). **Model R** predicts the box office earnings during the first theatrical week (forecasts are made prior to release) using 25 input variables. The values of Targets 1 and 2 are the same for Model R. **Model $\mathbf{W}_1$** predicts the box office earnings during the second theatrical week (Target 1) or up to the second theatrical week (Target 2) (forecasts are made one week after release) using 32 input variables, whereas **Model $\mathbf{W}_2$** predicts the box office earnings during the third theatrical week (Target 1) or up to the third theatrical week (Target 2) (forecasts are made two weeks after release) using 38 input variables. In step 3, significant and relevant variables are selected using the Akaike information criterion (AIC) (Hurvich & Tsai, 1989) for forecasting models based on screening data only. Because of the exponential increase in computational costs for the exhaustive search, we adopt a GA, which can find a pseudo-optimal solution efficiently as a variable selection method for the forecasting models that consider both screening and SNS data. In step 4, forecasting models are developed on the basis of four regression algorithms. In addition, the forecasting accuracy is enhanced by combining the forecasts made by individual models using the equally weighted average method as the combining rule.

[8] http://www.alias-i.com/lingpipe.

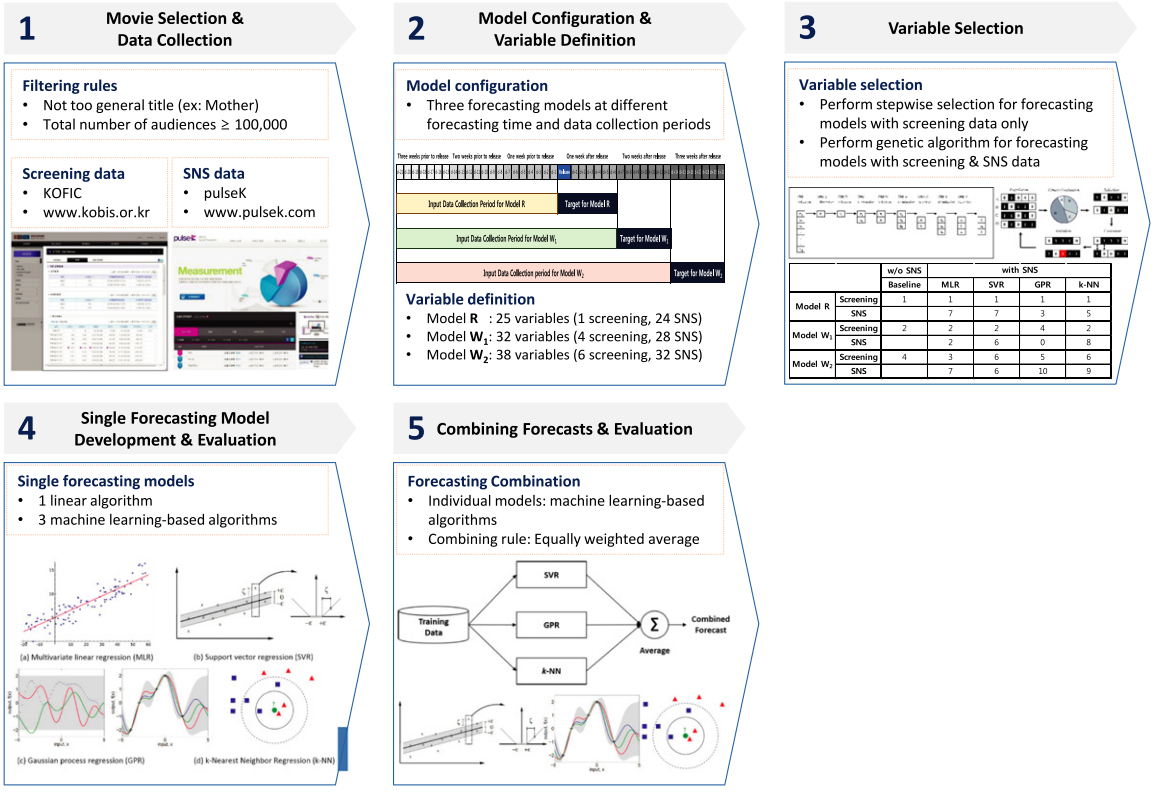


Fig. 1. The research framework for the development of box office forecasting models.

The prediction performances of the forecasting models developed are evaluated by employing a multivariate linear regression, using only screening-related variables as the baseline benchmark method. In order to substantiate the practical usability of the proposed forecasting models, we divided the entire dataset into validation and test datasets. The algorithm parameters are optimized using the validation dataset, while the forecasting performances are tested on the holdout dataset using the optimized parameters.

3.1. Motion picture selection and data collection

From among the motion pictures released in the Korean film market between Sep. 2011 and Dec. 2013, we selected 212 motion pictures based on the following selection criteria: (i) the total number of audiences is greater than 100,000 and (ii) the title is not too general (e.g., mother). The first rule was used for filtering out lower outliers that may cause the forecasting accuracy to degenerate (Chang & Ki, 2005; Dellarocas et al., 2007; Simonton, 2009; Wang, Cai, et al., 2010), whereas the second rule prevents irrelevant SNS mentions from being analyzed (Asur & Huberman, 2010). An example of such cases is where a mention contains the title words of a certain motion picture but is not genuinely referring to it (e.g., 2012, Mother, etc.). The number of motion pictures collected in this study can be considered sufficient in terms of the target market size and characteristics of the input variables. As is summarized in the Appendix, most studies devoted to box office forecasting have focused on the US film market, the largest in the world. The average and median numbers of motion pictures analyzed in previous studies are 361 and 76, respectively. When excluding the US domestic market, however, they decrease to 125 and 62, respectively. The number of motion pictures being analyzed in our study is the second largest in non-US film markets. In addition, the data collection period is restricted because we collect SNS mentions for the motion pictures; a sufficient volume of SNS data could be gathered only after the launch of major SNS sites (Facebook: 2004.02, Twitter: 2006.03). Some statistics of three target variables (cumulative box office earnings for the corresponding theatrical periods) are summarized in Table 2. As has been reported repeatedly in many previous studies (Calantone et al., 2010; Chintagunta et al., 2010; Elberse & Eliashberg, 2003; Lovallo et al., 2012; Marshall et al., 2013; Vany & Walls, 1999), all three target variables are positively skewed; a few blockbusters result in significantly large box office earnings, whereas the rest (the majority) ended up being mediocre movies. Thus, box office earnings are log-scaled when building forecasting models.

Two types of data were gathered from different sources. Screening-related data, such as the number of screens and seats on the day of release, weekly aggregated numbers of screens and seats, weekly numbers of audiences, and weekly box office earnings, were collected from the Korean Film Council.[9] In addition, SNS-related data, such as the weekly numbers of total, emotional, positive, and negative mentions in various SNSs, were collected from pulseK,[10]

[9] http://www.kofic.or.kr.

[10] http://www.pulsek.com.

Table 2
The average, standard deviation (Stdev.), minimum (Min.), maximum (Max.), and skewness of each target variable in KRW.

Model	Average	Stdev.	Min.	Max.	Skewness
Model R	4,847,348,551	5,070,423,315	244,505,000	29,246,291,000	2.1069
Model W$_1$	3,086,853,003	4,294,866,142	19,102,500	29,162,115,000	2.9099
Model W$_2$	1,527,466,244	2,425,119,891	10,000	15,640,387,500	2.5028

Table 3
Description of the motion picture-related attributes.

Num.	Category	Attribute	Description	p range
1	–	Index	Motion picture identifier	–
2 – 4		N_audience^p	Number of audiences in week p	1 – 3
5 – 7		Box_office^p	Box office in week p	1 – 3
8 – 10		N_screen$_p$	Number of screens at the beginning of week p	1 – 3
11 – 13		N_screen^p	Total number of screens in week p	1 – 3
14 – 16		N_seat$_p$	Number of seats at the beginning of week p	1 – 3
17 – 19		N_seat^p	Total number of seats in week p	1 – 3
35 – 39	SNS	N_mention^p	Total number of SNS mentions in week p	−3 – 2 except 0
40 – 44		N_emotional^p	Total number of emotional SNS mentions in week p	−3 – 2 except 0
45 – 49		N_positive^p	Total number of positive SNS mentions in week p	−3 – 2 except 0
50 – 54		N_negative^p	Total number of negative SNS mentions in week p	−3 – 2 except 0

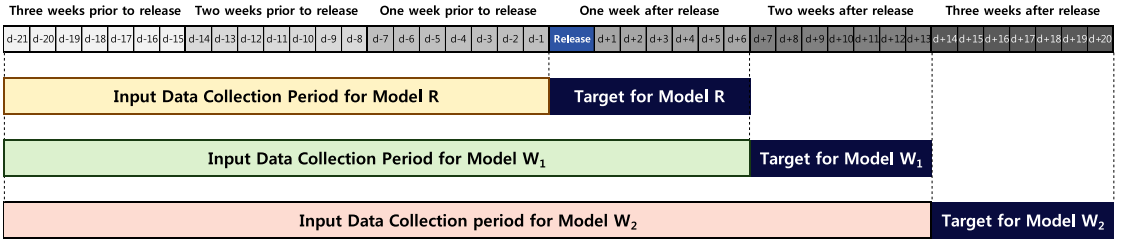


Fig. 2. The timing of the forecasting and data collection periods of each forecasting model.

which provides summarized SNS data after conducting a sentiment and polarity analysis. In this study, a total of 3726,902 mentions are collected; the movie that generated the most interest in cyberspace induced more than 160,000 mentions and blog posts over the 5-week period. In total, 53 attributes were considered initially in developing forecasting models. A brief description of each attribute is provided in Table 3.

3.2. Model configuration and variable definition

By definition, forecasting models are more valuable if the forecasts are made as early as possible. In other words, a forecast that is extraordinarily accurate is not useful if it is made only a few days before the closing. In order to maintain the utility of the forecasts, box office forecasting should be performed no later than two weeks from release. However, there is an information discrepancy between the periods prior to release and after release. When a motion picture has not yet been released, only a few pieces of screening-related information are available. All that we can know ahead of release are the numbers of reserved screens and seats on the day of release or the first week. On the other hand, once the motion picture has been released, additional screening-related data, such as weekly screens, seats, audiences, and weekly or cumulative box offices, become available.

In order to cope with the different degrees of data accessibility, three forecasting models are developed in this study, as is shown in Fig. 2. **Model R** forecasts the box office earnings in the first week (prior to release) on the basis of screening and SNS data collected over the three weeks prior to release. **Model W$_1$** forecasts the box office earnings during the second week (one week after release) on the basis of screening and SNS data collected over a four-week period (three weeks before release + one week after release), whereas **Model W$_2$** forecasts the box office earnings during the third week (two weeks after release) on the basis of screening and SNS data collected over a five-week period (three weeks before release + two week after release). The target variable and candidate input variables for each model are summarized in Table 4. Because there are four types of SNS mentions (i.e., total, emotional, positive, and negative) and SNS data collection periods differ according to the forecasting model, a total of 12, 16, and 20 original SNS-related variables are considered initially for **Model R**, **Model W$_1$**, and **Model W$_2$**, respectively. Then, 12 additional SNS-related variables are derived from the original variables: the cumulative/average increase/weekly increase of total/emotional/positive/negative mentions

Table 4
The candidate input variables for each forecasting model.

Category	Attribute	**Model R**	**Model W$_1$**	**Model W$_2$**
Target 1		Log$_{10}$(Box_office^1)	Log$_{10}$(Box_office^2)	Log$_{10}$(Box_office^3)
Target 2		Log$_{10}$(Box_office^1)	Log$_{10}$($\sum_{i=1}^{2}$Box_office^i)	Log$_{10}$($\sum_{i=1}^{3}$Box_office^i)
Screening	Original	N_seat$_1$	Box_office^1	Box_office^1+ Box_office^2
			N_seat$_2$	N_seat$_3$
			N_seat^1	N_seat^2
	Derived		Weekly_seat_increase	N_seat^1+ N_seat^2
				Weekly_seat_increase
				Weekly_cumulative_seat_increase
SNS	Original	N_mention^p	N_mention^p	N_mention^p
		N_emotional^p	N_emotional^p	N_emotional^p
		N_positive^p	N_positive^p	N_positive^p
		N_negative^p	N_negative^p	N_negative^p
		$p \in \{-3, -2, -1\}$	$p \in \{-3, -2, -1, 1\}$	$p \in \{-3, -2, -1, 1, 2\}$
	Derived	Cumulative_mention	Cumulative_mention	Cumulative_mention
		Cumulative_emotional	Cumulative_emotional	Cumulative_emotional
		Cumulative_positive	Cumulative_positive	Cumulative_positive
		Cumulative_negative	Cumulative_negative	Cumulative_negative
		Avg_mention_increase	Avg_mention_increase	Avg_mention_increase
		Weekly_mention_increase	Weekly_mention_increase	Weekly_mention_increase
		Avg_emotional_increase	Avg_emotional_increase	Avg_emotional_increase
		Weekly_emotional_increase	Weekly_emotional_increase	Weekly_emotional_increase
		Avg_positive_increase	Avg_positive_increase	Avg_positive_increase
		Weekly_positive_increase	Weekly_positive_increase	Weekly_positive_increase
		Avg_negative_increase	Avg_negative_increase	Avg_negative_increase
		Weekly_negative_increase	Weekly_negative_increase	Weekly_negative_increase

during each model's data collection period. Although the implicit concept of each variable name is identical across the models, their computations are different. For example, the average and weekly mention increases for **Model R**, **Model W$_1$**, and **Model W$_2$** are computed as follows:

$$\begin{aligned}
&\text{Avg\_mention\_increase } (\textbf{Model R}) \\
&\quad = \frac{\text{N\_mention}^{-1} - \text{N\_mention}^{-3}}{2}, \\
&\text{Weekly\_mention\_increase } (\textbf{Model R}) \\
&\quad = \text{N\_mention}^{-1} - \text{N\_mention}^{-2}, \\
&\text{Avg\_mention\_increase } (\textbf{Model W}_1) \\
&\quad = \frac{\text{N\_mention}^{1} - \text{N\_mention}^{-3}}{3}, \\
&\text{Weekly\_mention\_increase } (\textbf{Model W}_1) \\
&\quad = \text{N\_mention}^{2} - \text{N\_mention}^{1}, \\
&\text{Avg\_mention\_increase } (\textbf{Model W}_2) \\
&\quad = \frac{\text{N\_mention}^{2} - \text{N\_mention}^{-3}}{4}, \\
&\text{Weekly\_mention\_increase } (\textbf{Model W}_2) \\
&\quad = \text{N\_mention}^{2} - \text{N\_mention}^{1}. \qquad (1)
\end{aligned}$$

In addition to SNS variables, the three models also utilize different screening variables according to the data availability. For example, there is only one screening variable for **Model R**, i.e., the number of seats reserved for release day (N_seat$_1$). It should be noted that although the number of screens is also available, it is better to use the number of seats as the supply capacity for a certain motion picture because theatres are differently sized; some huge multiplexes may accommodate a larger audience, whereas some art theaters may accommodate under a hundred people. Once a motion picture is released, on the other hand, other screening variables, such as the cumulative box office and the cumulative number of seats in a certain week, are also available. For all three forecasting periods, the `Baseline` model, i.e., a multivariate linear regression using only screening-related variables, is also developed to validate the SNS-related variables and machine learning-based forecasting algorithms.

3.3. *Variable selection*

Because the number of candidate input variables for each model is quite large relative to the number of motion pictures analyzed, two variable selection techniques are adopted in this paper. Since the `Baseline` models have relatively few candidate input variables and adopt the linear regression as the forecasting algorithm, the best subset of input variables is determined based on the AIC. In this process, a multivariate linear regression is fitted for all possible input variable sets. If there are p variables, it is possible to construct $2^p - 1$ input variable sets. Note that, as there is only one screening-related variable for **Model R**, variable selection is not necessary. Hence, we only conduct the variable selection process for **Model W$_1$** and **Model W$_2$**. Since four and six screening-related variables are considered for **Model W$_1$** and **Model W$_2$**, respectively, a total of 15 and 63 candidate input variable sets are evaluated. For each input variable set, the AIC for the corresponding model is computed as follows:

$$AIC = n \log\left(\frac{SSE}{n}\right) + 2(p+1), \qquad (2)$$

where SSE, n, and p denote the sum of squared errors of the current model, the number of observations (the

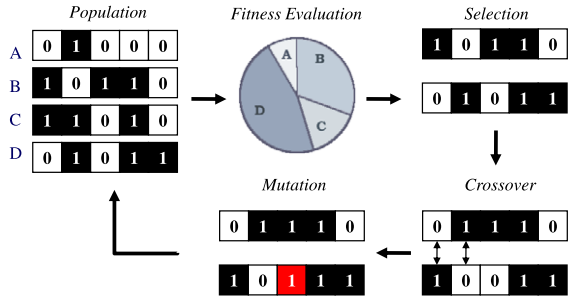


Fig. 3. GA for variable selection.

number of motion pictures in this study), and the number of predictors (input variables), respectively. The input variables of the model with the minimum AIC value are finally selected as the best subset.

Because the number of candidate input variables is relatively large when the SNS-related variables are considered in addition to the screening-related variables, there is a heavy computational cost involved in considering all possible subsets, and, in fact, it is practically impossible. For example, one has to evaluate 33,554,431 subsets for **Model R**, which has 25 candidate input variables. For the models that consider both SNS-related and screening-related variables, we employed a genetic algorithm (GA) based variable selection (Cho & Hermsmeier, 2002; Jarvis & Goodacre, 2004) as an alternative to AIC-based variable selection. Fig. 3 illustrates the variable selection procedure conducted by GA. GA searches the optimal set of input variables by mimicking natural evolutionary procedures such as selection, crossover, and mutation. Initially, a sufficient number of chromosomes, called a population, is created. Each chromosome has the form of a vector, having the same length of the total number of variables. Each cell in a chromosome, called a gene, has a value of either 1 or 0. The value 1 implies that the corresponding variable is activated during the modeling, whereas 0 implies deactivation. Forecasting models are trained on the basis of the variables that are activated by each chromosome, and their fitness values, i.e., the mean absolute percentage error (MAPE) in this paper, are calculated. Then, the chromosomes with higher fitness values are allowed to survive (selection) to produce the next generation. These surviving chromosomes exchange some part of their genes to produce new child chromosomes (crossover). Finally, the values of some randomly selected genes are changed with a very low probability (mutation) to create the opportunity to escape from a local optimum. Hence, by iterating the selection–crossover–mutation cycle a sufficient number of times, we can obtain a pseudo-optimal set of input variables.

3.4. Forecasting models

The following four regression algorithms are employed for building forecasting models, as shown in Fig. 4: `MLR`, `SVR`, `GPR`, and `k-NN`.

The `MLR` (Ross, 2004) fits the functional relationship between the input and target variables in the form of a linear equation (Fig. 4(a)), and has been used as the most fundamental forecasting algorithm in many applications (Goia, May, & Fusai, 2010; Jonsson, 1994; Zhang & Thomas, 2012). Although the `MLR` is analytically tractable, its performance is generally not as good as those of nonlinear algorithms. Let y_i denote the target value (total box office) of the ith motion picture, while x_{ij} denotes the jth input variable of the ith motion picture. Then, the `MLR` equation of d explanatory variables with n motion pictures can be written as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_d x_{id}, \quad \text{for } i = 1, 2, \ldots, n. \tag{3}$$

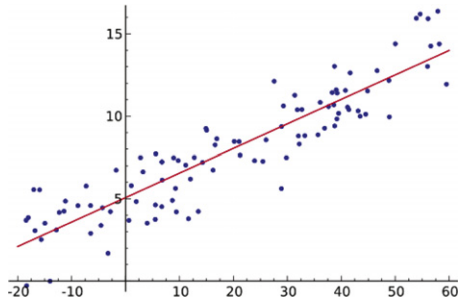
The regression coefficients $\boldsymbol{\beta}$ can be obtained by minimizing the squared error (residual) between the actual targets ($\mathbf{y}$) and the predictions made ($\hat{\mathbf{y}}$) using the ordinary least squares (OLS) method. The structure of the `MLR` used in this study is almost identical to one of the forecasting models introduced by Asur and Huberman (2010), in that the both adopt a multivariate linear regression as a regression algorithm and construct the input variable sets by considering the volume of SNS mentions in addition to the supply capacity.

The `SVR` (Smola & Schölkopf, 2004) is a nonlinear regression algorithm that is well known for its structural risk minimization (SRM) principle. The `SVR` fits the regression equation $\hat{y} = \mathbf{w}^T\mathbf{x} + b$, with the constraint of including as many training instances as possible in the ε-tube, as follows (Fig. 4(c)):

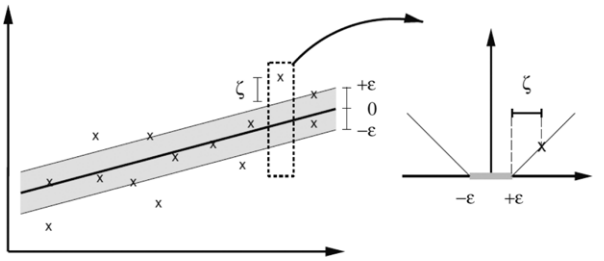
$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*) \\ \text{s.t} \quad & y_i - \mathbf{w}^T\mathbf{x}_i - b \le \varepsilon + \zeta_i, \quad \zeta_i \ge 0 \\ & \mathbf{w}^T\mathbf{x}_i + b - y_i \le \varepsilon + \zeta_i^*, \quad \zeta_i^* \ge 0. \end{aligned} \tag{4}$$

C in Eq. (4) controls the trade-off between the flatness and the error of the training samples outside the ε-tube. If C is set to a small value, a flatter regression function can be obtained but the observations will have larger errors. On the other hand, if C is set to a large value, the observations would have small errors, since the shape of the regression function becomes more complicated in order to encompass them within the ε-tube, but the regression function may lose the generalization ability due to the increased complexity. Therefore, the parameter C should be determined carefully so as to achieve both generality and forecasting accuracy. By eliminating the constraints using slack variables, the Primal Lagrangian formulation can be obtained:

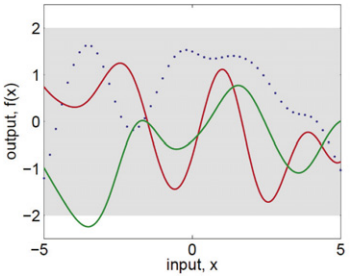
$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*) \\ & - \sum_{i=1}^{n}\alpha_i\left(\varepsilon + \zeta_i - y_i + \mathbf{w}^T\mathbf{x}_i + b\right) \\ & - \sum_{i=1}^{n}\eta_i\zeta_i - \sum_{i=1}^{n}\alpha_i^*\left(\varepsilon + \zeta_i^* + y_i - \mathbf{w}^T\mathbf{x}_i - b\right) \\ & - \sum_{i=1}^{n}\eta_i^*\zeta_i^* \\ \text{s.t} \quad & \alpha_i, \alpha_i^*, \zeta_i, \zeta_i^* \ge 0. \end{aligned} \tag{5}$$



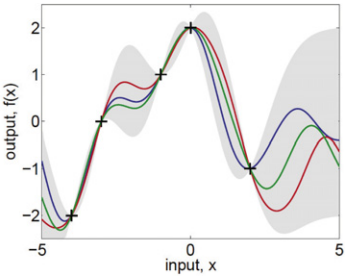
(a) Multivariate linear regression (MLR).

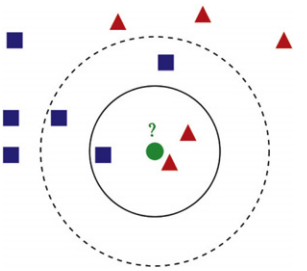


(b) Support vector regression (SVR).



(c) Gaussian process regression (GPR).





(d) k-nearest neighbor regression (k-NN).

Fig. 4. Four regression algorithms used for the forecasting models.

The optimal condition of the Lagrangian formulation above can be obtained by taking derivatives of the primal variables. Finally, Wolfe's dual problem is derived by replacing the condition in the primal problem as follows:

$$\begin{aligned} \max \quad & -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_i-\alpha_i^*)\mathbf{x}_i^T\mathbf{x}_j \\ & -\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*) \\ \text{s.t} \quad & \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0, \quad 0\le\alpha_i,\alpha_i^*\le 0. \end{aligned} \tag{6}$$

The SVR enables a nonlinear fitting by using a mapping $\phi(\mathbf{x})$ that transforms data from a low-dimensional input space into a high-dimensional feature space. Because only the inner products between input vectors are needed during the optimization process, a kernel trick $K(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j)$ is employed to compute the inner product in the feature space without an explicit mapping function.

The GPR (Rasmussen & Williams, 2005) begins with the Bayesian approach to the MLR and extends the expressiveness by adopting kernel tricks. In the GPR, the target y is expressed as a linear combination of the inputs with a Gaussian noise as follows:

$$y = f(\mathbf{x})+\epsilon, \quad f(\mathbf{x})=\mathbf{x}^T\mathbf{w}, \tag{7}$$

assuming that the noise follows an independent, identically distributed (i.i.d) Gaussian distribution with zero mean and variance σ^2. The likelihood, which is the probability density of the given data and parameters, can be obtained directly as

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\mathbf{w}) &= \prod_{i=1}^{n}p(y_i|\mathbf{x}_i,\mathbf{w}) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\Big(-\frac{1}{2\sigma^2}|\mathbf{y}-\mathbf{X}^T\mathbf{w}|^2\Big) \\ &= \mathcal{N}(\mathbf{X}^T\mathbf{w},\sigma^2 I). \end{aligned} \tag{8}$$

As the prior distribution over the weights, a zero mean Gaussian with covariance matrix $\mathbf{\Sigma}_p$ is generally used. Inference in the GPR is based on the posterior distribution over the weights by applying Bayes' theorem, and the prediction distribution for f_t at a test instance $\mathbf{x}_t$ is then obtained by averaging the output of all possible linear models with regard to the Gaussian posterior, as shown in Eq. (9):

$$\begin{aligned} p(f_t|\mathbf{x}_t,\mathbf{X},\mathbf{y}) &= \int f(\mathbf{x}_t|\mathbf{w})P(\mathbf{w}|\mathbf{X},\mathbf{y})d\mathbf{w} \\ &= \mathcal{N}\Big(\frac{1}{\sigma^2}\mathbf{x}_t^T\mathbf{A}^{-1}\mathbf{X}\mathbf{y},\mathbf{x}_t^T\mathbf{A}^{-1}\mathbf{x}_t\Big). \end{aligned} \tag{9}$$

As in the SVR, the GPR can fit a nonlinear relationship by introducing a set of basis functions $\phi(\mathbf{x})$ so as to project the data from a low dimensional space to a higher dimensional feature space, and using kernel tricks to compute dot products in the kernel space without an explicit form of $\phi(\mathbf{x})$.

The k-NN is the most widely employed instance-based learning algorithm. It is also called the memory-based reasoning or lazy learning method because, unlike other machine learning-based regression algorithms, it does not require an independent training procedure. The k-NN predicts the total box office of a new motion picture by combining the box offices of other motion pictures whose

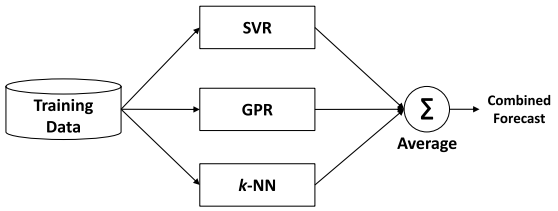


Fig. 5. Combination of three machine learning-based forecasting algorithms.

target values are already known and whose attribute values are similar. In order to do this, the k-NN first computes the similarity between the test observation and the reference observations using a pre-defined proximity measure. Then, the k most similar observations, called neighbors, are selected for the forecasting. The target value of the test observation is predicted by combining the target values of the selected neighbors as follows:

$$\hat{y}_i = \sum_{j \in k-NN(\mathbf{x}_i)} w_j y_j, \tag{10}$$

where $\hat{y}_i$, k-NN($\mathbf{x}_i$), w_j, and y_j denote the target value (total box office) of the ith motion picture, the index set of k nearest neighbors of $\mathbf{x}_i$, the weight assigned to the neighbor $\mathbf{x}_j$, and the target value of the neighbor $\mathbf{x}_j$, respectively. There are two questions with the k-NN. (1) How many neighbors should be selected? (2) How should the weights be assigned? To address these questions, we adopted the locally linear reconstruction algorithm proposed by Kang and Cho (2008).

In addition to the four individual regression models, another forecast is also made based on the combination of three machine learning-based regression algorithms, to enhance the forecasting accuracy. Fig. 5 illustrates the way in which the individual forecasts are combined. First, each of the machine learning-based algorithms, namely SVR, GPR, and k-NN, is trained using the same training data. Next, their forecasts are combined via an aggregation method, to generate a single forecast of the combination model. As an outcome aggregation method, equally weighted averaging is adopted.

3.5. Validation method and performance measure

In order to validate our proposed forecasting models, we conduct a sequential validation and test experiment. First, the entire data set is divided into two subsets; the newest movies, those released over the most recent three months, are set aside as a holdout dataset for the independent test, while the remaining movies are used for parameter optimization. Forty-three movies belong to the test dataset, while 169 movies belong to the validation dataset.

Because the machine learning-based forecasting models have their own parameters to be determined (i.e., the types of kernels and their specifications, together with the cost for the errors, in the SVR, the kernel width and hyperparameters for the prior distributions in the GPR, and the number of nearest neighbors in the k-NN), 10-fold cross-validation was used for the selection of the best parameter. In a 10-fold cross-validation, the validation dataset is divided into 10 subgroups, and each subgroup in turn is set aside for validation, whereas the remaining nine subgroups are used for training the forecasting model. Under each parameter setting, this validation is repeated for all 10 folds, and their aggregated forecasting performance is recorded. Once the validation processes have been completed for all parameter combinations, the most accurate parameter combination is finally selected for further analysis.

The test dataset, which has never been used before, is used to evaluate the forecasting performance of each model for unseen data. In this test experiment, each forecasting algorithm is trained based on the entire validation dataset with its best parameter combination, which is determined in the 10-fold cross-validation. Then, it produces a forecast for every movie in the test dataset.

Once the forecasts have been produced by each model, they are evaluated in terms of the MAPE (Abel et al., 2010; Ainslie et al., 2005; Lee et al., 2012; Marshall et al., 2013; Sawhney & Eliashberg, 1996; Wang, Zhang, et al., 2010) and the root mean squared error (RMSE) (Jun et al., 2011; Simonoff & Sparraw, 2000) as performance measures:

$$\begin{aligned} \text{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \end{aligned} \tag{11}$$

where y_i and $\hat{y}_i$ denote the actual and predicted box office revenues of the motion picture i. We should note that the adjusted R^2 is not used as a performance measure in this study, even though it is the most commonly adopted performance measure for box office forecasting models (Asur & Huberman, 2010; Brewer et al., 2009; Chintagunta et al., 2010; Eliashberg & Shugan, 1997; Wang, Cai, et al., 2010). The reason why the adjusted R^2 has become the *de facto* performance measure in box office forecasting is that most of the forecasting models are rooted in the multivariate linear regression algorithm. The adjusted R^2, which captures the ratio of the amount of the variance explained by the regression model to the total variance, is appropriate only when the input and target variables are related linearly. This study neither assumes linearity between the input and target variables nor develops a linear model alone. Thus, more general performance criteria that can be used for both linear and nonlinear regression algorithms should be adopted.

4. Results

4.1. Variable selection

Tables 5–7 summarize the variable selection results for each forecasting algorithm in each forecasting period. Note that the Baseline is the multivariate linear regression with the screening variables only, and its variables are selected by the AIC from among all possible cases. In contrast, the other four algorithms consider both the screening and the SNS variables; in these cases, the

Table 5
The variables selected for each forecasting algorithm for **Model R**.

Category	Variable	Baseline	MLR	SVR	GPR	k-NN	Total
Screening	N_seat_1	Y	Y	Y	Y	Y	5
SNS	$N\_mention^{-3}$	–	Y	Y		Y	3
	$N\_mention^{-2}$	–	Y			Y	2
	$N\_mention^{-1}$	–					0
	$N\_emotional^{-3}$	–					0
	$N\_emotional^{-2}$	–		Y			1
	$N\_emotional^{-1}$	–	Y	Y	Y		3
	$N\_positive^{-3}$	–				Y	1
	$N\_positive^{-2}$	–					0
	$N\_positive^{-1}$	–	Y	Y			2
	$N\_negative^{-3}$	–					0
	$N\_negative^{-2}$	–			Y		1
	$N\_negative^{-1}$	–				Y	1
	Cumulative_mention	–				Y	1
	Cumulative_emotional	–	Y				1
	Cumulative_positive	–					0
	Cumulative_negative	–					0
	Avg_mention_increase	–	Y				1
	Avg_emotional_increase	–					0
	Avg_positive_increase	–	Y	Y			2
	Avg_negative_increase	–					0
	Weekly_mention_increase	–	Y	Y			2
	Weekly_emotional_increase	–			Y		1
	Weekly_positive_increase	–					0
	Weekly_negative_increase	–					0
Total number of variables		1	8	8	4	6	

Table 6
The variables selected for each forecasting algorithm for **Model W_1**.

Category	Variable	Baseline	MLR	SVR	GPR	k-NN	Total
Screening	$Log_{10}(Box\_office^1)$	Y	Y	Y	Y	Y	5
	N_seat_2				Y		1
	N_seat^1				Y		1
	Weekly_seat_increase	Y	Y	Y	Y	Y	5
SNS	$N\_mention^{-3}$	–					0
	$N\_mention^{-2}$	–					0
	$N\_mention^{-1}$	–					0
	$N\_mention^1$	–					0
	$N\_emotional^{-3}$	–				Y	1
	$N\_emotional^{-2}$	–					0
	$N\_emotional^{-1}$	–					0
	$N\_emotional^1$	–		Y		Y	2
	$N\_positive^{-3}$	–					0
	$N\_positive^{-2}$	–					0
	$N\_positive^{-1}$	–		Y		Y	2
	$N\_positive^1$	–					0
	$N\_negative^{-3}$	–					0
	$N\_negative^{-2}$	–					0
	$N\_negative^{-1}$	–		Y		Y	2
	$N\_negative^1$	–		Y		Y	2
	Cumulative_mention	–				Y	0
	Cumulative_emotional	–					0
	Cumulative_positive	–	Y	Y			2
	Cumulative_negative	–					0
	Avg_mention_increase	–					0
	Avg_emotional_increase	–					0
	Avg_positive_increase	–	Y				1
	Avg_negative_increase	–				Y	1
	Weekly_mention_increase	–				Y	1
	Weekly_emotional_increase	–					0
	Weekly_positive_increase	–					0
	Weekly_negative_increase	–		Y		Y	2
Total number of variables		2	4	8	4	10	

Table 7
The variables selected for each forecasting algorithm for **Model W_2**.

Category	Variable	Baseline	MLR	SVR	GPR	k-NN	Total
Screening	Log_{10} (Box_office^1 + Box_office^2)	Y	Y	Y	Y	Y	5
	N_seat$_3$	Y		Y	Y	Y	4
	N_seat^2			Y	Y	Y	3
	N_seat^1 + N_seat ^{2}			Y		Y	2
	Weekly_seat_increase	Y	Y	Y	Y	Y	5
	Weekly_cumulative_seat_increase	Y	Y	Y	Y	Y	5
SNS	N_mention^{-3}	–			Y		1
	N_mention^{-2}	–					0
	N_mention^{-1}	–					0
	N_mention^1	–					0
	N_mention^2	–		Y	Y		2
	N_emotional^{-3}	–	Y				1
	N_emotional^{-2}	–	Y				1
	N_emotional^{-1}	–					0
	N_emotional^1	–	Y				1
	N_emotional^2	–			Y	Y	2
	N_positive^{-3}	–		Y	Y	Y	3
	N_positive^{-2}	–					0
	N_positive^{-1}	–					0
	N_positive^1	–					0
	N_positive^2	–			Y	Y	2
	N_negative^{-3}	–					0
	N_negative^{-2}	–					0
	N_negative^{-1}	–					0
	N_negative^1	–	Y			Y	2
	N_negative^2	–					0
	Cumulative_mention	–				Y	0
	Cumulative_emotional	–		Y	Y		2
	Cumulative_positive	–	Y	Y	Y		3
	Cumulative_negative	–				Y	1
	Avg_mention_increase	–	Y	Y			2
	Avg_emotional_increase	–					0
	Avg_positive_increase	–					0
	Avg_negative_increase	–	Y		Y		2
	Weekly_mention_increase	–				Y	1
	Weekly_emotional_increase	–			Y	Y	2
	Weekly_positive_increase	–			Y	Y	2
	Weekly_negative_increase	–		Y		Y	2
	Total number of variables	4	10	12	15	15	

variables are selected by GA. For **Model R**, four to eight variables are included, depending on the forecasting algorithms; the only screening variable, i.e., N_seat$_1$, is selected for all algorithms, while only two original variables, N_mention^{-3} and N_emotional^{-1}, are selected for more than three algorithms from the SNS variables. In addition, 10 variables are never selected by any of the forecasting models. An interesting observation for **Model R** is the proportion of derived variables in each algorithm. Four derived variables, such as Cumulative_emotional, Avg_mention_increase, Avg_positive_increase, and Weekly_mention_increase, are selected for the MLR, which is the only linear algorithm, and they occupy half of the selected variable set. For the nonlinear algorithms, on the other hand, only one (GPR and k-NN) or two (SVR) derived variables are selected, and their portions are below 25%. With regard to the characteristics of mentions, the volumes of informative and sentimental mentions are almost equally important in box office forecasting prior to release. We therefore select nine variables that represent the total volume of mentions and 14 emotion-related variables for **Model R**.

For **Model W_1**, two screening variables, Box_office^1 and Weekly_seat_increase, are selected for the Baseline algorithm; these variables are also selected by all four algorithms when the SNS variables are taken into consideration as well. In addition, the other two screening variables are also selected for the GPR. It should also be noted that the GPR is built based on screening variables only, although SNS-related variables are taken into account simultaneously. For the other forecasting algorithms, on the other hand, at least two SNS-related variables are identified as significant. Unlike the GPR, eight SNS-related variables are required to build a forecasting model when the k-NN algorithm is employed. Among the selected variables, all but two of the variables (Cumulative_mention and Weekly_mention_increase in the k-NN) are related to emotional feelings. This indicates that once the motion picture is released, as distinct from in **Model R**, the volume of mentions expressing one's sentiment becomes more critical than the volume of informative mentions for box office forecasting for motion pictures.

For **Model W_2**, the distinctive difference between its variable selection results and those of **Model R** and **Model W_1** is the higher proportion of variables selected.

Table 8
Forecasting performance of each algorithm for Target 1 (non-cumulative box office earnings) in each model for the holdout dataset, in terms of MAPE.

Model	Linear algorithm		Machine learning				Combination
	Baseline	MLR	SVR	GPR	k-NN	Average	
Model R	0.8626	0.5958 (30.93%)	0.4278 (50.40%)* (28.19%)$^{+}$	0.5087 (41.02%)* (14.62%)$^{+}$	0.4605 (46.61%)* (22.71%)$^{++}$	0.4657 (46.01%) (21.84%)	0.4496 (47.88%)* (24.54%)$^{++}$ (3.45%)
Model W_1	0.8877	0.8927 (−0.56%)	0.6102 (31.27%)* (31.65%)$^{+}$	0.5421 (38.93%)** (39.28%)$^{++}$	0.6499 (26.80%) (27.21%)$^{+}$	0.6007 (32.33%) (32.71%)	0.5335 (39.90%)** (40.24%) (11.19%)
Model W_2	1.0409	0.8218 (21.05%)	0.5444 (47.70%)** (33.76%)$^{++}$	0.5671 (45.52%)** (31.00%)$^{+}$	0.6182 (40.61%)* (24.77%)$^{+}$	0.5765 (44.61%) (29.84%)	0.5319 (48.90%)** (35.27%)$^{++}$ (7.74%)

Notes: The number in the column titled "Average" is the average MAPE of SVR, GPR, and k-NN.
* Indicates that the forecasting performance of the corresponding algorithm is better than the Baseline at the 0.2 significance level.
** Indicates that the forecasting performance of the corresponding algorithm is better than the Baseline at the 0.1 significance level.
*** Indicates that the forecasting performance of the corresponding algorithm is better than the Baseline at the 0.05 significance level.
$^{+}$ Indicates that the forecasting performance of the algorithm is better than the MLR at the 0.2 significance level.
$^{++}$ Indicates that the forecasting performance of the algorithm is better than the MLR at the 0.1 significance level.
$^{+++}$ Indicates that the forecasting performance of the algorithm is better than the MLR at the 0.05 significance level.

For **Model W_2**, 34% of the variables are identified as significant on average, whereas 25% of the SNS variables are found to be significant. For **Model R** and **Model W_1**, the ratios are 26% and 20% for all variables and 22% and 14% for the SNS variables, respectively. Similarly to the variable selection result of **Model W_1**, the cumulative box office and Weekly_seat_increase are selected by all forecasting algorithms, but in this case, Weekly_cumulative_seat_increase is also selected by all forecasting algorithms. Note also that all screening variables are found to be significant for the SVR and the k-NN. Among the SNS variables, it seems that positive mentions play a more significant role than other types of mentions in forecasting the box office. Two positive mention-related variables, N_positive^{-3} and Cumulative_positive, are identified as being significant for more than three algorithms. In addition, positive mention-related variables survive 10 times, whereas total, emotional, and negative-related variables survive 8, 8, and 7 times, respectively.

Based on the variable selection results for the three forecasting models, it can be concluded that the role of SNS-related variables changes with the forecasting time shift. When the forecast is made prior to release, the total volume of mentions is a key determinant; since no one has actually watched the motion picture at that time, the public's expectations of a certain movie are indicated by the frequency of SNS mentions. Once the movie has been released, on the other hand, the volume of emotional mentions becomes a key determinant, and the significance of the informative mentions gradually disappears. These results indicate that the volume of positive mentions is a key indicator of a movie's success in the long term.

4.2. Forecasting accuracy

The proposed forecasting models are verified by decomposing the forecasting accuracy improvement into three stages according to the usage of the SNS data, machine learning algorithms, and combining forecasts. As was demonstrated in Section 3.2, the Baseline is a multivariate linear regression that uses only the screening-related data. In order to investigate the predictive power of the SNS data without the help of sophisticated machine learning-based algorithms, the forecasting accuracy of a multivariate linear regression using both screening and SNS data (MLR) is computed. The second and third improvements are explored by comparing the accuracies of individual machine learning-based algorithms (SVR, GPR, and k-NN) and a combination of the three machine learning-based algorithms (Combination), respectively.

Tables 8 and 9 show the forecasting performances of the different forecasting algorithms in terms of MAPE for Targets 1 and 2, respectively, while Tables 10 and 11 do the same in terms of the RMSE. Note that the MAPE is computed based on the original box office earnings, whereas the RMSE is computed based on the logged box office earnings. The structures of the four tables are identical. The numbers in the column titled "Average" are the average MAPE/RMSE values of SVR, GPR, and k-NN for the same forecasting time. The numbers in the first set of parentheses in the columns MLR, SVR, GPR, k-NN, "Average", and Combination denote the performance improvement made by each algorithm relative to the Baseline, whereas the numbers in the second set of parentheses in the columns SVR, GPR, k-NN, "Average", and Combination denote the performance improvement relative to MLR. Further, the numbers in the third set of parentheses in the Combination column denote the performance improvement relative to the "Average". In order to indicate the statistical significance of these improvements, paired t-tests are conducted, with ***, **, and * indicating that the forecasting performance of the corresponding algorithm is better than that of the Baseline at the 0.05, 0.1 and 0.2 significance levels, respectively. $^{+++}$, $^{++}$, and $^{+}$ denote that the forecasting performance of the algorithm is better than that of the MLR at the 0.05, 0.1, and 0.2 significance levels, respectively.

Table 9
Forecasting performance of each algorithm for Target 2 (cumulative box office earnings) in each model for the holdout dataset, in terms of MAPE.

Model	Linear algorithm		Machine learning				Combination
	Baseline	MLR	SVR	GPR	k-NN	Average	
Model R	0.8626	0.5958 (30.93%)	0.4278 (50.40%)* (28.19%)$^{+}$	0.5087 (41.02%)* (14.62%)$^{+}$	0.4605 (46.61%)* (22.71%)$^{++}$	0.4657 (46.01%) (21.84%)	0.4496 (47.88%)* (24.54%)$^{++}$ (3.45%)
Model W$_1$	0.1664	0.1535 (7.76%)	0.1251 (24.83%) (18.51%)	0.1206 (27.52%) (21.43%)	0.1288 (22.62%) (16.11%)	0.1248 (24.99%) (18.68%)	0.1216 (26.93%) (20.78%) (2.58%)
Model W$_2$	0.0624	0.0587 (5.90%)	0.0473 (24.16%)* (19.41%)$^{+}$	0.0541 (13.29%) (7.86%)	0.0486 (22.15%) (17.27%)	0.0500 (19.87%) (14.84%)	0.0453 (27.38%)* (22.83%)$^{+}$ (9.38%)

Note: See the notes to Table 8.

Table 10
Forecasting performance of each algorithm for Target 1 (non-cumulative box office earnings) in each model for the holdout dataset, in terms of RMSE.

Model	Linear algorithm		Machine learning				Combination
	Baseline	MLR	SVR	GPR	k-NN	Average	
Model R	0.2897	0.2868 (1.00%)	0.2623 (9.46%)* (8.54%)$^{+}$	0.2634 (9.09%)* (8.17%)$^{+}$	0.2635 (9.05%)* (8.13%)$^{+}$	0.2631 (9.20%) (8.28%)	0.2549 (12.02%)* (11.13%)$^{+}$ (3.11%)
Model W$_1$	0.3088	0.3096 (−0.28%)	0.2666 (13.66%)** (13.91%)$^{++}$	0.2850 (7.71%) (7.97%)	0.2625 (14.97%)** (15.21%)$^{++}$	0.2714 (12.11%) (12.36%)	0.2394 (22.46%)*** (22.68%)$^{+++}$ (11.78%)
Model W$_2$	0.3817	0.3607 (5.52%)	0.3090 (19.05%)*** (14.32%)$^{++}$	0.3398 (11.00%)* (5.80%)	0.3622 (5.12%) (−0.43%)	0.3370 (11.72%) (6.57%)	0.3094 (18.95%)** (14.21%)$^{++}$ (8.19%)

Note: See the notes to Table 8.

Table 11
Forecasting performance of each algorithm for Target 2 (cumulative box office earnings) in each model for the holdout dataset, in terms of RMSE.

Model	Linear algorithm		Machine learning				Combination
	Baseline	MLR	SVR	GPR	k-NN	Average	
Model R	0.2897	0.2868 (1.00%)	0.2623 (9.46%)* (8.54%)$^{+}$	0.2634 (9.09%)* (8.17%)$^{+}$	0.2635 (9.05%)* (8.13%)$^{+}$	0.2631 (9.20%) (8.28%)	0.2549 (12.02%)* (11.13%)$^{+}$ (3.11%)
Model W$_1$	0.0756	0.0766 (−1.28%)	0.0663 (12.34%)** (13.45%)	0.0606 (19.95%)*** (20.97%)$^{+++}$	0.0669 (11.60%) (12.72%)	0.0646 (14.63%) (15.71%)	0.0616 (18.57%)** (19.60%) (4.61%)
Model W$_2$	0.0347	0.0345 (0.59%)	0.0339 (2.26%)* (1.68%)$^{++}$	0.0373 (−7.40%) (−8.04%)	0.0306 (11.79%) (11.27%)	0.0339 (2.22%) (1.64%)	0.0314 (9.49%)** (8.95%)$^{+++}$ (7.44%)

Note: See the notes to Table 8.

When only the screening-related variables are used, the forecasting error rates seem to be relatively high; the MAPE of Targets 1 and 2 is 86.26% for **Model R**. In terms of the RMSE, the error rates of the Baseline are 0.2897 for the holdout dataset. This means that when no historical box office records are available, the forecasting error rate in the first theatrical week exceeds 85% of the actual box office earnings (MAPE), and the average box office difference is about 1.7 billion KRW (RMSE), provided that the average box office earning for **Model R** is about 4.85 billion KRW. An interesting observation is that, as additional screening-related information becomes available after release, the forecasting error rates increase slightly for Target 1 (non-cumulative box office earnings), whereas they decrease considerably for Target 2 (cumulative box office earnings); the MAPE increases by up to 100% in **Model W$_2$** for Target

1, whereas it drops to less than 7% in **Model W**$_2$ for Target 2. This is caused mainly by a general trend in weekly box office earnings and the existence of failed movies. Since the actual box office earnings are highest in the first week and decrease steadily as time goes by, the denominator of the MAPE also decreases steadily for Target 1, while increasing for Target 2. Thus, the same sized forecasting error makes a larger contribution to the MAPE of Target 1 than to that of Target 2. With regard to the existence of failed movies, for example, one motion picture in the holdout dataset attracted only 10 audiences (10,000 KRW) in the third week. Since the forecasting models cannot adapt to these extraordinary outliers, the residuals for those motion pictures are usually very large for Target 1 (non-cumulative box office earnings), which leads to a significant degradation in the overall forecasting performance. As a result, for Target 1, the MAPEs for **Model W**$_1$ and **Model W**$_2$ increase by 2.51%p and 17.83%p, respectively, compared to those of **Model R**, while the RMSEs also increase by 0.0191 and 0.092, respectively. When it comes to Target 2, on the other hand, since the cumulative box office earnings are used as the target variable, the effect of extremes can be smoothed out, resulting in residuals which fluctuate less. Since the cumulative box office earnings up to the current week are included as an input variable, which is highly correlated with the cumulative box office earnings up to the next week, the predictive power of the Baseline for Target 2 improves substantially. The MAPEs are reduced to 16.64% and 6.24% for **Model W**$_1$ and **Model W**$_2$, respectively, and similar trends can be found in terms of RMSE. Compared to **Model R**, the RMSE decreases by 73.90% and 88.02% for **Model W**$_1$ and **Model W**$_2$, respectively.

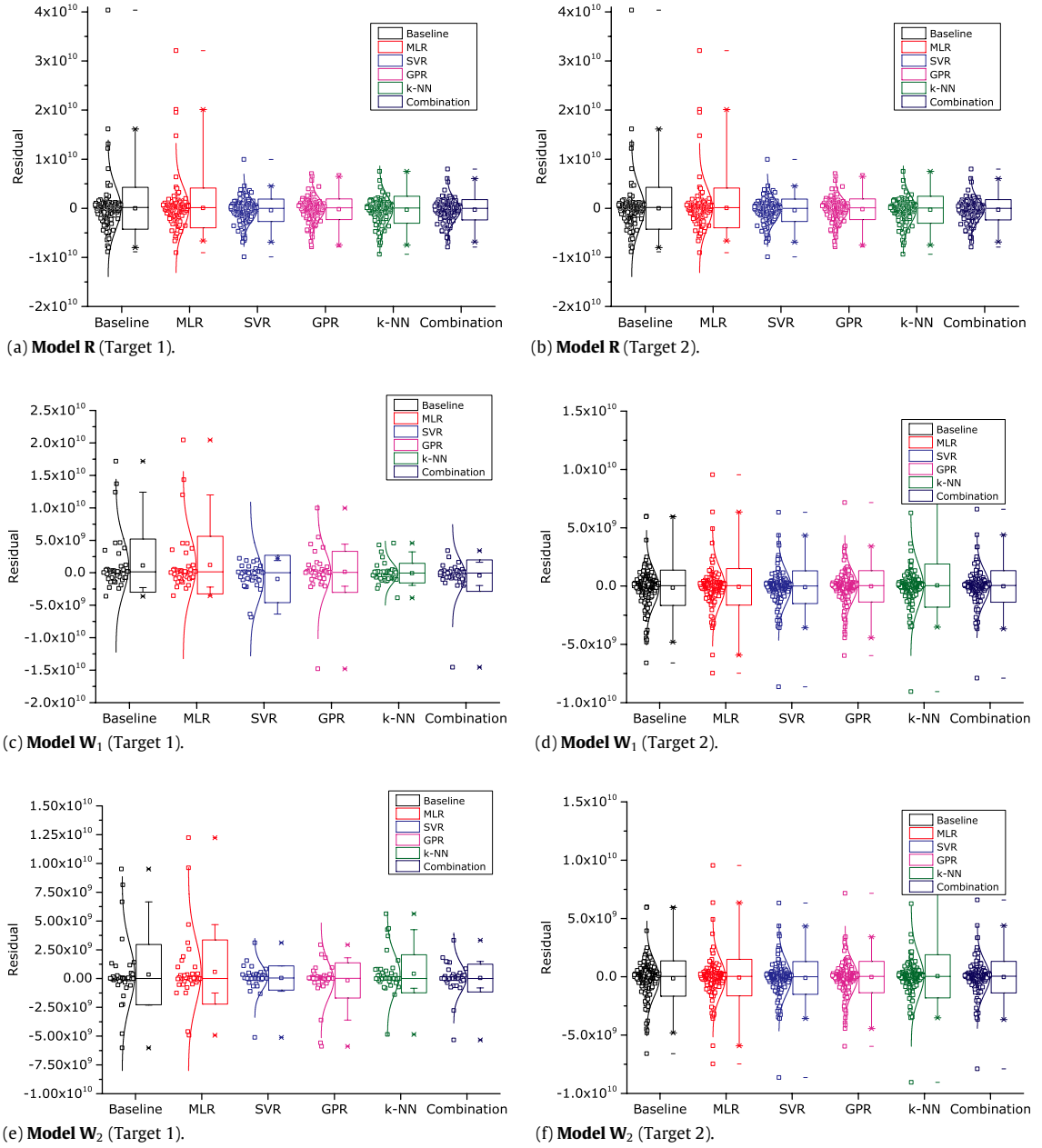
It is found that the inclusion of SNS variables makes a remarkable improvement to the forecasting accuracies of all three forecasting periods, and the effect of this inclusion is especially outstanding when forecasts are made prior to release. The MAPE of the MLR decreases to 0.5958, which is 30.94% lower than the Baseline in **Model R**. In terms of the RMSE, utilizing SNS data improves the forecasting performance by 1.00%. It should be pointed out that the RMSE improvement is smaller than that of the MAPE because the MAPE measures the relative difference between actual box office earnings and forecast earnings, whereas the RMSE measures the absolute difference between them. Therefore, the reduced residual errors (actual box office earnings − box office forecast) for movies with either very low or very high actual box office earnings contribute equally to the improvement in the RMSE, while the contribution to the improvement in the MAPE is magnified by reducing the residuals for movies with relatively low actual box office earnings. From a managerial perspective, the MAPE would be a more suitable indicator as long as financial success is typically measured by the return on investment (ROI), which is the ratio of the actual box office earnings to the total amount of the investment.

In addition to the SNS variables, the adoption of machine learning-based forecasting algorithms also makes a noticeable improvement to the forecasting accuracy. In all three forecasting periods, all three machine learning-based algorithms result in lower MAPEs and RMSEs than the MLR for both Target 1 (non-cumulative box office earnings) and Target 2 (cumulative box office earnings), with the exception of one case (GPR in **Model W**$_2$ for the Target 2 in terms of RMSE), and many of these performance improvements are found to be statistically significant. For **Model R**, machine learning-based algorithms improve the forecasting accuracy by 46.01% in terms of the MAPE, while the RMSE decreases by 9.20% compared to the Baseline. In addition, most of the improvements resulting from machine learning-based algorithms are confirmed to be statistically significant relative to both the Baseline and the MLR for **Model R**. For **Model W**$_1$ and **Model W**$_2$, machine learning-based algorithms enhance the forecasting performance by at least 25% (Target 1) and 13% (Target 2) compared to the Baseline. In comparison with the MLR, machine learning-based algorithms reduce the MAPE by more than 14% (Target 1) and 7% (Target 2) on average for all cases. It is hard to conclude which machine learning-based algorithm is superior. The best algorithm varies depending on the forecasting periods and datasets. The SVR is found to be the best in seven out of 12 cases (three forecasting periods × two targets × two performance measures), followed by the GPR for three cases and the k-NN for two cases. However, it is worth noting that the performance differences between any pairs of machine learning-based algorithms are not statistically significant.

Finally, our study also finds that combining the forecasts of machine learning-based algorithms can boost the forecasting performance too. Not only are the MAPE/RMSE of Combination the lowest among the forecasting algorithms, its improvements are also statistically significant in most cases. For Target 1, the improvements in the MAPE from Combination relative to the average of the three machine learning-based algorithms are 3.45%, 11.19%, and 7.74% for **Model R**, **Model W**$_1$, and **Model W**$_2$, respectively. For Target 2, the corresponding values are 3.45%, 2.58%, and 9.38%, respectively. Similar trends can be found in terms of the RMSE. For both Targets 1 and 2, the improvements in the RMSE from Combination are larger than 3%, 4%, and 7%, respectively. Interestingly, for Target 2, the effect of combining different algorithms is positively correlated with the average forecasting performances of individual algorithms, in terms of both the MAPE and the RMSE. Despite the fact that the forecasting precisions of machine learning-based algorithms are lowest in **Model R** and highest in **Model W**$_2$, the Combination gains the greatest relative improvement for **Model W**$_2$.

Summarizing the results of Tables 8–11 helps us to verify that all three of the remedies proposed in this paper, i.e., utilizing SNS data, employing machine leaning-based nonlinear algorithms, and combining individual forecasts, undoubtedly improve the forecasting performance. Each remedy, however, makes a different contribution to the improvement; using SNS data has a greater impact than the machine learning-based algorithms in early forecasting (**Model R**), but the influence becomes equivalent or even reversed in later forecasting. Combining forecasts enhances the forecasting performance further, even if the individual algorithms have already secured a considerable forecasting accuracy.

Fig. 6 shows the scatter and box plots of the residuals in each stage of the three forecasting models. According to



(a) **Model R** (Target 1). (b) **Model R** (Target 2).

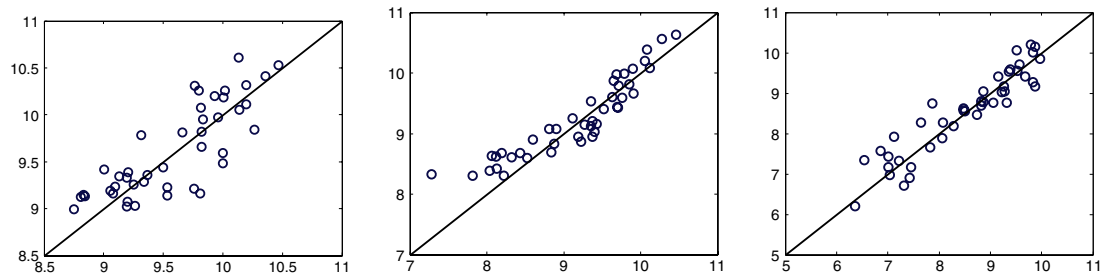
(c) **Model W**$_1$ (Target 1). (d) **Model W**$_1$ (Target 2).

(e) **Model W**$_2$ (Target 1). (f) **Model W**$_2$ (Target 2).

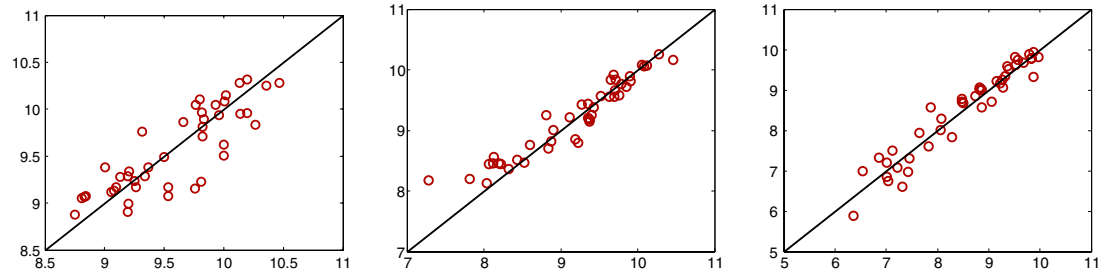
Fig. 6. Scatter and box plots of the residuals ($y - \hat{y}$) for the three forecasting models with the non-cumulative box office earnings (left column) and the cumulative box office earnings (right column).

this figure, not only does the `Combination` result in the lowest average forecasting error, its individual forecasts are also preferable. In the case of **Model R** (Fig. 6(a) and (b)), it is evident that the distribution of the residuals produced by the `Combination` is narrowed, and the 1–99 quantile range is reduced compared to other individual algorithms. In the cases of **Model W**$_1$ and **Model W**$_2$ (Fig. 6(c), (d), (e), and (f)), the distribution of the residuals and the 1–99 quantile range are not easily distinguishable with the individual machine learning-based algorithms, but it is still recognizable that the upper and lower outliers of the `Combination` are closer to zero than those of either the linear algorithms or individual machine learning-based algorithms.

Figs. 7 and 8 exhibit the relationship between the box office forecasts of the `Baseline` or the `Combination` (y-axis) and the actual box office earnings (x-axis) for each forecasting period for Targets 1 and 2, respectively. In the case of **Model R**, it is worth noting that the `Baseline` has a critical problem: it usually overestimates the box office earnings when the actual values fall under one of the extremes (Fig. 7(a) and Fig. 8(a)). When the actual box office earnings are either lower than 1 billion KRW (9 on the x-axis on the logged scale) or higher than 15.8

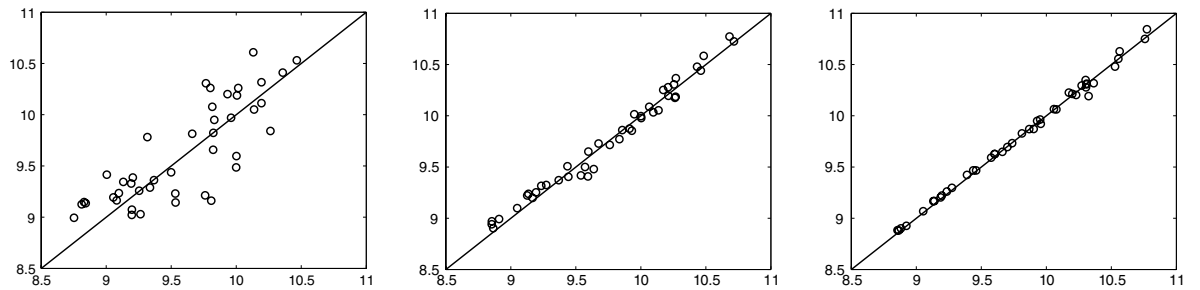


(a) **Model R** (`Baseline`). (b) **Model W**$_1$ (`Baseline`). (c) **Model W**$_2$ (`Baseline`).

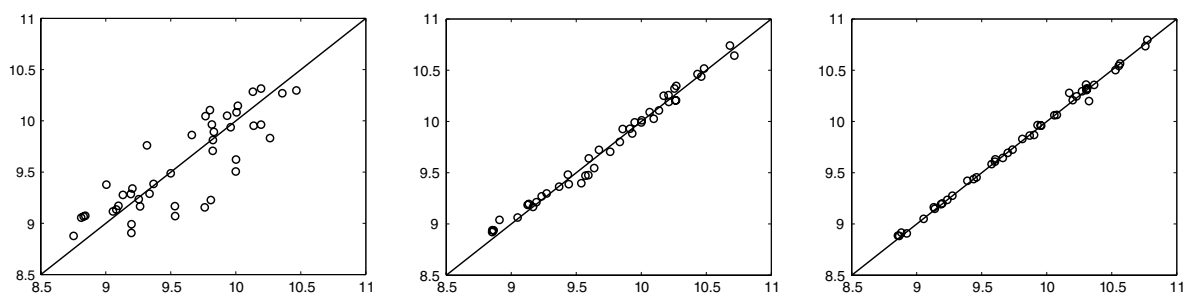


(d) **Model R** (`Combination`). (e) **Model W**$_1$ (`Combination`). (f) **Model W**$_2$ (`Combination`).

Fig. 7. Box-office forecasts (y-axis) against the actual box office earnings (x-axis) of the `Baseline` and `Combination` for each forecasting period (Target 1; both the actual box office earnings and the forecasts are log-transformed).



(a) **Model R** (`Baseline`). (b) **Model W**$_1$ (`Baseline`). (c) **Model W**$_2$ (`Baseline`).



(d) **Model R** (`Combination`). (e) **Model W**$_1$ (`Combination`). (f) **Model W**$_2$ (`Combination`).

Fig. 8. Box-office forecasts (y-axis) against the actual box office earnings (x-axis) of the `Baseline` and `Combination` for each forecasting period (Target 2; both the actual box office earnings and the forecasts are log-transformed).

billion KRW (10.2 on the x-axis on the logged scale), most of the points are located above the equivalence line. On the other hand, the `Combination` apparently resolves the overestimation problem for these blockbuster movies (Fig. 7(d) and Fig. 8(d)). In the case of **Model W**$_1$ and **Model W**$_2$, it is not as easy to discover the critical flaws of `Baseline`, since it also produces quite accurate forecasts, owing to the inclusion of the historical box office earnings in the input variable set. However, it is nevertheless true that the points are distributed more closely around

the equivalence line for the Combination than for the Baseline forecasts.

On the basis of the experimental results presented in Tables 8–11 and Figs. 6–8, we can draw the following conclusions. First, as time passes, the forecasting accuracy can be improved without any treatment. The MAPEs of the Baseline algorithm are over 85% prior to release, but drop to below 6.5% two weeks after release when the cumulative box office earnings are considered. In terms of RMSE, the forecasting accuracies are higher than 0.28 when forecasting is done prior to release, but they decrease substantially to less than 0.04 for the cumulative box office earnings (Target 2) two weeks after release. Second, the use of SNS data does improve the forecasting accuracy. It is the most remarkable factor for enhancing the accuracy, particularly when forecasting is carried out prior to release. Once historical box office earnings are available, the contribution of the SNS data to the forecasting performance is not as impressive as it was prior to release, but it certainly drives additional performance improvements. Third, the use of machine learning-based algorithms is effective in forecasting over all of the forecasting periods. Merely employing machine learning-based nonlinear algorithms improves the MAPE by at least 14% (Target 1) and 7% (Target 2), and in conjunction with the SNS data it improves the MAPE by more than 26% (Target 1) and 20% (Target 2). Fourth, a combination of the forecasts of individual machine learning-based algorithms can boost the predictive power, irrespective of the forecasting period. Finally, the collaboration of SNS data, machine learning-based algorithms, and the forecast combination enhances the performance by more than 40% and 26% in terms of MAPE, for Targets 1 and 2, respectively, compared with the conventional model.

5. Conclusion

In this paper, a new approach to forecasting the box office earnings of motion pictures is proposed. Three consecutive forecasting models are developed by adapting current screening and SNS information in order to make the forecasts more accurate. For each model, two types of target variables are defined: box office earnings for the first, second, and third weeks (Target 1), and the cumulative box office earnings up to the first, second, and third weeks (Target 2). In order to eliminate the subjectivity of the variable configuration, only screening and WOM-related information was used. In contrast to other studies, the audience's preference for a certain motion picture was measured by analyzing ordinary SNS usage behaviors, rather than by voluntary and explicitly provided user ratings. In order to enhance the forecasting accuracy, a forecast combination of the machine learning-based regression algorithms was employed. Algorithm parameters for individual forecasting models were determined based on the 10-fold validation using 169 motion pictures, and the forecasting performances were tested on the holdout dataset, which consisted of 43 motion pictures and had not been used elsewhere, so as to evaluate the practical effectiveness of the proposed approach. Over three different forecasting periods, the proposed model improved the forecasting accuracy by more than 40% (Target 1) and 26% (Target 2) in terms of MAPE relative to the linear regression on the basis of only screening-related variables.

Apart from the notable experimental results, the current study has a few limitations that suggest further research directions. First, we assume that the number of seats is determined solely by exhibitors. However, in practice, the box office earnings is a primary factor considered by exhibitors when reallocating limited screens for the upcoming days or weeks. Thus, there exists an endogeneity problem, i.e., the forecast of one factor becomes the predictor of another factor, between the supply capacity for a certain movie and its future box office earnings (Elberse & Eliashberg, 2003; Neelamegham & Chintagunta, 1999; Sawhney & Eliashberg, 1996). Since this endogeneity problem is critical from a managerial perspective, it would be worthwhile investigating whether and how SNS data and machine learning algorithms can resolve it. Second, most of the variables, irrespective of whether they are related to movie characteristics or WOM, only capture the information on the target movie. However, one of the main factors that should be considered in box office forecasting is the competition environment surrounding each movie. For example, if a motion picture is released simultaneously with other blockbusters, it may attract smaller audiences than it could have had but for the competing blockbusters. Therefore, dealing with the competition environment, ranging from competitor identification to the quantification of their influence, would be a natural extension of the current study. Third, the experiment could suffer from a sampling bias because only 212 motion pictures released to the Korean market are taken into account, due to SNS data availability issues. A more general conclusion could be reached if motion pictures from a wide range of markets were taken into consideration. Fourth, because the forecasting models aim to support exhibitors in their managerial decisions on screen reallocation, their practical effects should be analyzed by applying the forecasting models to real situations or by conducting carefully designed simulations.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, and Future Planning (NRF-2014R1A1A1004648).

Appendix A. Survey on input variables in box office forecasting

See Table A.1.

Appendix B. Summary of motion picture data collection

See Table B.1.

Table A.1
Survey of the descriptions and significance of input variables in box office forecasting.

Category	Variable	Research	Description	Significance
Movie-specific	Screens	Asur and Huberman (2010)	N. of theaters (daily)	N/A
		Calantone et al. (2010)	N. of screens (weekly)	○
		Chintagunta et al. (2010)	Total N. of theaters where the movie is playing simultaneously	○
		Elberse and Eliashberg (2003)	N. of screens (weekly)	○
		Gong et al. (2011)	N. of theaters at a film's widest point of distribution	○
		Liu (2006)	N. of screens (weekly)	○
		Lovallo et al. (2012)	N. of theaters in the opening week	○
		Neelamegham and Chintagunta (1999)	N. of screens (weekly)	○
		Qin (2011)	N. of screens (daily)	○
		Ravid (1999)	N. of screens at release	○
		Simonoff and Sparraw (2000)	N. of screens for the first weekend	○
		Wang, Cai, et al. (2010)	N. of screens (opening day)	○
		Wen and Yang (2011)	N. of copies	○
		Zhang and Skiena (2009)	N. of screens at release	○
		Zufryden (1996)	N. of theaters (weekly)	○
	Age	Chintagunta et al. (2010)	Days on screen	○
		Neelamegham and Chintagunta (1999)	Weeks on screen	○
		Qin (2011)	Days on screen	○
		Wang, Zhang, et al. (2010)	Weeks on screen	○
		Zufryden (1996)	Weeks on screen	○
	Star power	Brewer et al. (2009)	Number of individuals from the master list who appear in a given film	×
		Elberse and Eliashberg (2003)	1–100 scale scored based on the Hollywood Reporter Star Power Index	○
		Gong et al. (2011)	Leading actor star power, based on cumulative box office performance of starred films (0–100 scale)	○
		Jun et al. (2011)	0 (no stars), 1 (famous actors/actresses), 2 (superstars)	×
		Litman (1983)	Whether it contained any top 10 box office stars for previous two years	×
		Lovallo et al. (2012)	Whether two lead actors were listed in the "Top Ten Money-Making Stars" list any of the previous three years	×
		Neelamegham and Chintagunta (1999)	Dummy variable (major star possessing marque value)	△
		Ravid (1999)	Dummy for unknown lead actors (references to lead actors in any guide)	×
			Whether the movie includes actors or directors who had won Academy awards	×
			Whether the movie includes actors who had top-grossing movies in the previous year	×
		Sawhney and Eliashberg (1996)	Dummy variable (major star possessing marque value)	△
		Simonoff and Sparraw (2000)	Number of actors (actresses) listed in Entertainment Weekly's 25 Best Actors (Actresses)	○
			Number of actors listed in Top 20 box office per movie in their career list according to the Movie Times Web	○
		Wen and Yang (2011)	Quantity variable (0–10) with the professional critic rating on websites or databases	×
	Advertising	Chintagunta et al. (2010)	Average per-day advertising until the movie is released in a new market	○
		Dellarocas et al. (2007)	Estimated pre-release marketing cost	○
		Elberse and Eliashberg (2003)	Advertising expenditures	○
		Eliashberg et al. (2000)	Proportion of respondents intending to see the movie after exposure to movie advertising	N/A
		Gong et al. (2011)	Marketing cost	○
		Jun et al. (2011)	Advertising expenditure	×
		Lovallo et al. (2012)	Number of theaters in the opening week (proxy)	○
		Wang, Zhang, et al. (2010)	Volume of media publicity (weekly)	○
		Wen and Yang (2011)	Promotion cost	○
	Award	Lee (2009)	Number of drama/non-drama award nominations	△
		Litman (1983)	Dummies for Academy winners/nominees	○
		Simonoff and Sparraw (2000)	Dummies for Academy winners/nominees	○
		Wen and Yang (2011)	Dummies for nominees	×

(continued on next page)

Table A.1 (*continued*)

Category	Variable	Research	Description	Significance
Movie-specific	Budget	Brewer et al. (2009)	Production budget	◯
		Gong et al. (2011)	Production cost	◯
		Lee (2009)	Production budget	◯
		Litman (1983)	Production cost	◯
		Lovallo et al. (2012)	Production budget	×
		Ravid (1999)	Production cost	◯
		Simonoff and Sparraw (2000)	Production budget	×
		Wen and Yang (2011)	Production budget	×
		Zhang et al. (2009)	Production cost	◯
	Critic	Brewer et al. (2009)	Rotten Tomato rating in percentage	◯
		Dellarocas et al. (2007)	Average professional critics reviews	◯
		Elberse and Eliashberg (2003)	1–5 scale scored based on Entertainment weekly (assigned by leading newspaper critics)	◯
		Eliashberg and Shugan (1997)	N. of 'pro', 'con', 'mixed' reviews from Variety magazine	△
		Litman (1983)	N. of stars in the rating of New York Daily News	◯
		Liu (2006)	N. of critical reviews (weekly), percentage of positive reviews	◯
		Ravid (1999)	Ratio of good reviews	×
		Sawhney and Eliashberg (1996)	Standardized professional star evaluations	△
		Simonoff and Sparraw (2000)	Rating from Roger Ebert (well known film critic in Chicago Sun-Times) in 0–4 stars	×
		Wen and Yang (2011)	Quantity variable (0–10) with the specialist ratings	N/A
	Director	Elberse and Eliashberg (2003)	1–100 scale scored based on the Hollywood Reporter Star Power Index	×
		Gong et al. (2011)	Director star power based on cumulative box office performance of directed films (0–100 scale)	◯
		Wen and Yang (2011)	Quantity variable (0–10) with the professional critic rating on websites or databases	×
	Distributor	Calantone et al. (2010)	Dummy variable (major distributor)	×
		Gong et al. (2011)	Dummy variable (major distributor)	◯
		Jun et al. (2011)	Dummy variable (major distributor)	×
		Litman (1983)	Dummy variable (major distributor)	×
		Neelamegham and Chintagunta (1999)	Dummy variable (independent distributor)	△
		Wen and Yang (2011)	Sum of the works of the distributor release	×
	Genre	Brewer et al. (2009)	Movie genre (dummy variables for each)	◯
		Calantone et al. (2010)	Movie genre (dummy variables for each)	△
		Gong et al. (2011)	Dummy variable for hi-tech genre	×
		Lee (2009)	Movie genre (dummy variables for comedy and adventure)	×
		Litman (1983)	Movie genre (dummy variables for each)	×
		Lovallo et al. (2012)	Movie genre (dummy variables for each)	×
		Neelamegham and Chintagunta (1999)	Movie genre (dummy variables for each)	△
		Qin (2011)	Movie genre	◯
		Sawhney and Eliashberg (1996)	Movie genre (dummy variables for each)	△
		Simonoff and Sparraw (2000)	Movie genre (dummy variables for each)	◯
		Wen and Yang (2011)	Movie genre (dummy variables for each)	×
	Income	Brewer et al. (2009)	US personal income for the month of release	◯
	Nationality	Simonoff and Sparraw (2000)	Classified as US, English-speaking countries, and non-English speaking countries	×
		Wang, Zhang, et al. (2010)	Dummy variable (Chinese = 1)	◯
		Zhang and Skiena (2009)	Dummy variable (US and others)	◯
	Price	Brewer et al. (2009)	Ticket prices for theatrical plays, movies and concerts for the month of release	×
		Wen and Yang (2011)	Ticket price	◯
	Rating	Brewer et al. (2009)	MPAA ratings	×
		Dellarocas et al. (2007)	MPAA ratings	◯
		Gong et al. (2011)	MPAA ratings	◯
		Litman (1983)	MPAA ratings	×
		Qin (2011)	MPAA ratings	×
		Ravid (1999)	MPAA ratings	◯
		Sawhney and Eliashberg (1996)	MPAA ratings	◯
		Simonoff and Sparraw (2000)	MPAA ratings	◯
		Zhang and Skiena (2009)	MPAA ratings	◯

(continued on next page)

Table A.1 (*continued*)

Category	Variable	Research	Description	Significance
Movie-specific	Sequel	Brewer et al. (2009)	The gross of the film preceding the sequel	◯
		Gong et al. (2011)	Comparative analysis between sequels and others	◯
		Lovallo et al. (2012)	Dummy variable (sequel = 1)	◯
		Ravid (1999)	Dummy variable (sequel = 1)	◯
		Sawhney and Eliashberg (1996)	Dummy variable (sequel = 1)	△
		Simonoff and Sparraw (2000)	Dummy variable (sequel = 1)	×
		Wen and Yang (2011)	Dummy variable (sequel = 1)	×
		Zhang and Skiena (2009)	Dummy variable	×
	Timing of release	Brewer et al. (2009)	Binary dummy variables (released in May–July, or during Thanksgiving and Christmas holidays)	◯
		Duan et al. (2008)	Dummy variable (weekend: Fri. to Sun.)	◯
		Elberse and Eliashberg (2003)	1–100 scale scored based on AC Nielsen EDI, weekly basis	×
		Gong et al. (2011)	Dummies for summer (May–August) and holiday (November–December)	◯
		Litman (1983)	Dummies for holiday (November–December), Easter (March–April), and summer (June–August)	△
		Qin (2011)	Dummy variable	◯
		Ravid (1999)	Continuous measure for seasonality (1 for Christmas and 0.35 for early December)	×
			Dummy for Christmas season (December 18–31)	×
		Simonoff and Sparraw (2000)	Dummy for holiday weekend (President's Day, Memorial Day, Labor Day, Thanksgiving, or the Christmas season)	×
			Dummy for summer season (Memorial Day–Labor Day)	◯
		Wang, Zhang, et al. (2010)	Dummy variable (New Year holidays)	◯
		Wen and Yang (2011)	Nominal variable (1–6)	×
		Zhang and Skiena (2009)	Dummy variable for holiday season	◯
	Competition	Calantone et al. (2010)	N. of incumbents (older than 2 weeks) with same genre	◯
			N. of incumbents (older than 2 weeks) with different genre	×
			N. of newly entrants with same/different genre	◯
		Chintagunta et al. (2010)	Average critic score of competing movies in the new market (0–100)	×
			Average star power of competing movies in the new market (0–6)	×
		Elberse and Eliashberg (2003)	N. of similar movies (same genre or ratings) in Top 25 divided by their ages (weeks)	◯
		Gong et al. (2011)	Rank of the film on the opening weekend (0–3)	×
		Liu (2006)	N. of new released movies, average age of movies (both top 20)	×
		Wen and Yang (2011)	Whether there is a power substitute movie released the same period	×
WOM	Awareness	Asur and Huberman (2010)	N. of tweets per hours	◯
		Calantone et al. (2010)	Average revenues per screen (weekly)	◯
		Chintagunta et al. (2010)	Cumulative volume of reviews until the movie is released in a new market	×
		Dellarocas et al. (2007)	N. of user ratings	◯
		Duan et al. (2008)	N. of user reviews (daily)	◯
		Eliashberg et al. (2000)	Average frequency of word-of-mouth conversation (simulated)	N/A
		Liu (2006)	N. of messages on blogs (weekly)	◯
		Neelamegham and Chintagunta (1999)	Cumulative viewership (proxy)	△
		Qin (2011)	N. of blog posts (daily)	◯
		Wang, Cai, et al. (2010)	Incidence on Yahoo! movies	◯
		Wang, Zhang, et al. (2010)	N. of SNS posts (weekly)	◯
	Preference	Asur and Huberman (2010)	Emotional tweets over neutral tweets, positive tweets over negative tweets	◯
		Chintagunta et al. (2010)	Mean rating of reviews until the movie is released in a new market	◯
		Duan et al. (2008)	Average user grades (daily)	×
		Eliashberg et al. (2000)	Proportion of respondents intending to see movie after exposure to positive (negative) WOM	N/A
		Liu (2006)	Percentages of positive and negative blog messages (weekly)	×
		Wen and Yang (2011)	Audience rating (0–10)	×

◯, △, and × denote that the corresponding variable is determined to be certainly significant, partially significant, and not significant, respectively.

Table B.1
Markets, numbers of motion pictures, data collection periods, and movie selection criteria used in previous studies.

Research	Market	N. movies	Periods	Criteria
Litman (1983)	US	125	1972–1978	Production cost information is available
Eliashberg and Sawhney (1994)	US	1	1989	(1) Not be too popular or well known, (2) should generate sufficient variability in enjoyment ratings, (3) not a stereotypical thriller, horror movie, or romantic comedy
Sawhney and Eliashberg (1996)	US	19	1990–1991	N/A
Zufryden (1996)	France (US film)	63	1993.01–1993.06	N/A
Eliashberg and Shugan (1997)	US	36	1991–1992	Complete box office life is available
Jedidi et al. (1998)	US	102	1990.12–1992.4	(1) The top 5 movies for at least one week, (2) wide-release and mass-marketing
Krider and Weinberg (1998)	US	2	1992	N/A
Vany and Walls (1999)	US	2015	1984–1996	N/A
Ravid (1999)	US	175	1991–1993	Eliminating all very low budget films
Neelamegham and Chintagunta (1999)	US and 13 other countries	35	1994.01–1996.05	Based on the Variety magazine sample (top 60 in US and top 10 in international markets)
Simonoff and Sparraw (2000)	US	311	1998	Relevant business information is available
Eliashberg et al. (2000)	Netherlands (US film)	2	1993	N/A
Elberse and Eliashberg (2003)	US, 4 European countries	164	1999	(1) Produced or co-produced in the US, (2) released in the US in 1999, (3) appeared at least once in the US box office top 25
Ainslie et al. (2005)	US	825	1995.06–1998.06	All movies released domestically
Liu (2006)	US	40	2002.05–2002.09	Excluding titles for which there is no reliable information about all control variables
Delen et al. (2007)	US	849	1998–2002	N/A
Duan et al. (2008)	US	71	2003–2004	Based on Variety's 2003–2004 box office rank in the U.S. market, (2) user review data are available
Brewer et al. (2009)	US	466	1997–2001	Top 100 domestic grossing films per year
Zhang and Skiena (2009)	US	1500	1960–2008	N/A
Lee (2009)	9 East Asian markets	585	2002–2007	Top 100 US-produced movies in the US market
Zhang et al. (2009)	China	241	2005–2006	Cut the sample set on an average 40 movies for a class (6 classes for revenue size)
Lee and Chang (2009)	Korea	100	2005.01–2006.10	N/A
Abel et al. (2010)	US	23	2008.10–2008.12	Contain over 100 million blog posts
Asur and Huberman (2010)	US	24	2009.11–2010.02	(1) Released on Friday, (2) in wide release, (3) not too general title (e.g., 2012)
Calantone et al. (2010)	US	2948	1997–2004	N/A
Chakravartya et al. (2010)	US	1	2004	Based on the pretest with 60 undergraduate students, the subjects exhibited neither extreme interest nor extreme disinterest
Chintagunta et al. (2010)	US	148	2003.11–2005.02	User ratings data is available
Wang, Cai, et al. (2010)	US	66	2006–2007	(1) Released by the major studios, (2) excluding movies that were not released nationwide, not released in theaters, and contained missing variables
Wang, Zhang, et al. (2010)	China	51	2006.10–2009.03	(1) Movies released by major Chinese studios, (2) excluding movies with missing values
Gong et al. (2011)	US	2016	2002–2007	Omitting films without corresponding production and marketing cost data
Qin (2011)	US	49	2009.03–2009.10	Top 13 box office based on their opening weekend revenue
Jun et al. (2011)	11 countries	43	1997	N/A
Wen and Yang (2011)	China	60	2005–2010	Top 10–15 box office
Lovallo et al. (2012)	US	19	2005	Scheduled to be released in the summer season
Mestyán et al. (2013)	US	312	2010	Track the corresponding page in Wikipedia
Lee et al. (2012)	Korea	40	2005.12–2010.09	Half of successful movies (more than 2M viewers), half of unsuccessful movies, randomly selected
Marshall et al. (2013)	Chile	117	2001–2003	Played for three weeks or longer

References

Abel, F., Diaz-Aviles, E., Henze, N., Krause, D., & Siehndel, P. (2010). Analyzing the blogosphere for predicting the success of music and movie products. In *Proceedings of the 2010 international conference on advances in social networks analysis and mining* (pp. 276–280). Denmark: Odense.

Ainslie, A., Drèze, X., & Zufryden, F. (2005). Modeling movie life cycles and market share. *Marketing Science*, *24*(3), 508–517.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology - Volume 01* (pp. 492–499). Washington, DC, USA: IEEE Computer Society, URL http://dx.doi.org/10.1109/wi-iat.2010.63.

Brewer, S. M., Kelley, J. M., & Jozefowicz, J. J. (2009). A blueprint for success in the US film industry. *Applied Economics*, *41*(5), 589–606.

Calantone, R. J., Yeniyurt, S., Townsend, J. D., & Schmidt, J. B. (2010). The effects of competition in short product life-cycle markets: the case of motion pictures. *The Journal of Product Innovation Management*, *27*(3), 349–361.

Chakravartya, A., Liub, Y., & Mazumdarc, T. (2010). The differential effects of online word-of-mouth and critics' reviews on pre-release movie evaluation. *Journal of Interactive Marketing*, *24*(3), 185–197.

Chang, B.-H., & Ki, E.-J. (2005). Devising a practical model for predicting theatrical movie success: focusing on the experience good property. *Journal of Media Economics*, *18*(4), 247–269.

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: accounting for sequential rollout and aggregation across local markets. *Marketing Science*, *29*(5), 944–957.

Cho, S. J., & Hermsmeier, M. A. (2002). Genetic algorithm guided selection: variable selection and subset selection. *Journal of Chemical Information and Modeling*, *42*(4), 927–936.

Delen, D., Sharda, R., & Kumar, P. (2007). Movie forecast guru: a web-based DSS for Hollywood managers. *Decision Support Systems*, *43*(4), 1151–1170.

Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: the case study of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23–45.

Duan, W., Gub, B., & Whinston, A. B. (2008). The dynamics of online word-of-mouth and product sales an empirical investigation of the movie industry. *Journal of Retailing*, *84*(2), 233–242.

Elberse, A., & Eliashberg, J. (2003). Demand and supply dynamics for sequentially released products in international markets: the case of motion pictures. *Marketing Science*, *22*(3), 329–354.

Eliashberg, J., Jonker, J.-J., Sawhney, M. S., & Wierenga, B. (2000). MOVIEMOD: an implementable decision-support system for prerelease market evaluation of motion pictures. *Marketing Science*, *19*(3), 226–243.

Eliashberg, J., & Sawhney, M. S. (1994). Modeling goes to Hollywood: predicting individual differences in movie enjoyment. *Management Science*, *40*(9), 1151–1173.

Eliashberg, J., & Shugan, S. M. (1997). Influencers or predictors? *Journal of Marketing*, *61*(2), 68–78.

Goia, A., May, C., & Fusai, G. (2010). Functional clustering and linear regression for peak load forecasting. *International Journal of Forecasting*, *26*(4), 700–711.

Gong, J. J., Young, S. M., & der Stede, W. A. V. (2011). Real options in the motion picture industry: Evidence from film marketing and sequels. *Contemporary Accounting Research*, *28*(5), 1438–1466.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307.

Jarvis, R. M., & Goodacre, R. (2004). Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *Bioinformatics*, *21*(7), 860–868.

Jedidi, K., Krider, R. E., & Weinberg, C. B. (1998). Clustering at the movies. *Marketing Letters*, *9*(4), 393–405.

Jonsson, B. (1994). Prediction with a linear regression model and errors in a regressor. *International Journal of Forecasting*, *10*(4), 549–555.

Jun, D. B., Kim, D. S., & Kim, J. H. (2011). *A Bayesian DYMIMIC model for forecasting movie viewers*. KAIST business school working paper series (KCB-WP-2011-003).
URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1972062.

Kang, P., & Cho, S. (2008). Locally linear reconstruction for instance-based learning. *Pattern Recognition*, *41*(11), 3507–3518.

Krider, R. E., & Weinberg, C. B. (1998). Competitive dynamics and the introduction of new products: the motion picture timing game. *Journal of Marketing Research*, *35*(1), 1–15.

Lee, F. L. F. (2009). Cultural discount of cinematic achievement: the academy awards and U.S. movies' east Asian box office. *Journal of Cultural Economics*, *33*(4), 239–263.

Lee, K. J., & Chang, W. (2009). Bayesian belief network for box-office performance: a case study on Korean movies. *Expert Systems with Applications*, *36*(1), 280–291.

Lee, Y., Kim, S.-H., & Cha, K. C. (2012). A generalized Bass model for predicting the sales patterns of motion pictures having seasonality and herd behavior. *Journal of Global Scholars of Marketing Science: Bridging Asia and the World*, *22*(4), 310–326.

Litman, B. R. (1983). Predicting success of theatrical movies: an empirical study. *Journal of Popular Culture*, *16*(4), 159–175.

Liu, Y. (2006). Word of mouth for movies: its dynamics and impact on box office revenue. *Journal of Marketing*, *70*(3), 74–89.

Lovallo, D., Clarke, C., & Camerer, C. (2012). Robust analogizing and the outside view: two empirical tests of case-based decision making. *Strategic Management Journal*, *33*(5), 496–512.

Marshall, P., Dockendorff, M., & Ibáñez, S. (2013). A forecasting system for movie attendance. *Journal of Business Research*, *66*(10), 1800–1806.

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PLoS ONE*, *8*(8), e71226.

Neelamegham, R., & Chintagunta, P. (1999). A Bayesian model to forecast new product performance in domestic and international markets. *Marketing Science*, *18*(2), 115–136.

Qin, L. (2011). Word-of-blog for movies: a predictor and an outcome of box office revenue? *Journal of Electronic Commerce Research*, *12*(3), 187–198.

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian processes for machine learning (adaptive computation and machine learning)*. The MIT Press.

Ravid, S. A. (1999). Information, blockbusters, and stars: a study of the film industry. *The Journal of Business*, *72*(4), 463–492.

Rogers, E. (1976). New product adoption and diffusion. *Journal of Consumer Research*, *2*(2), 290–301.

Ross, S. M. (2004). *Introduction to probability and statistics for engineers and scientists*. San Diego, CA, USA: Academic Press.

Sawhney, M. S., & Eliashberg, J. (1996). A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science*, *15*(2), 113–131.

Simonoff, J. S., & Sparraw, I. R. (2000). Predicting movie grosses: winners and losers, blockbusters and sleepers. *Chance*, *13*(3), 15–24.

Simonton, D. K. (2009). Cinematic success criteria and their predictors: the art and business of the film industry. *Psycology and Marketing*, *26*(5), 400–420.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

Vany, A. D., & Walls, W. D. (1999). Uncertainty in the movie industry: does star power reduce the terror of the box office? *Journal of Cultural Economics*, *23*(4), 285–318.

Wang, F., Cai, R., & Huang, M. (2010). Forecasting movie-going behavior based on online pre-release wom and opening strength. *2nd international workshop on intelligent systems and applications* (pp. 1–4).

Wang, F., Zhang, Y., Li, X., & Zhu, H. (2010). Why do moviegoers go to the theater? The role of prerelease media publicity and online word of mouth in driving moviegoing behavior. *Journal of Interactive Advertising*, *11*(1), 50–62.

Wen, K., & Yang, C. (2011). Determinants of the box office performance of motion picture in China – indication for Chinese motion picture market by adapting determinants of the box office (part II). *Journal of Science and Innovation*, *1*(4), 17–26.

Zhang, L., Luo, J., & Yang, S. (2009). Forecasting box office revenue of movies with BP neural network. *Expert Systems with Applications*, *36*(3), 6580–6587.

Zhang, W., & Skiena, S. (2009). Improving movie gross prediction through news analysis. *IEEE/WIC/ACM international joint conferences on web intelligence and intelligent agent technologies, WI-IAT'09* (Vol. 1, pp. 301–304).

Zhang, J., & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting*, *28*(1), 204–215.

Zufryden, F. S. (1996). Linking advertising to box office performance of new film releases: a marketing planning model. *Journal of Advertising Research*, *36*(4), 29–42.

Taegu Kim received his BS, MS, and Ph.D, all in Industrial Engineering, from Seoul National University in 2002, 2004, and 2013, respectively. His research interests include the mathematical modeling of innovation diffusion, data mining, and investor behavior analysis based on econometrics.

Jungsik Hong received his BS, MS, and Ph.D, all in Industrial Engineering, from Seoul National University in 1982, 1985, and 1988. Currently, he is a Professor of Industrial and Information Systems Engineering at Seoul National University of Technology. He is a member of the Korea Institute of Industrial Engineers and the Korean Operations Research Society, and has worked as a referee of *IIE Transaction*. His research interests include the mathematical modeling of innovation diffusion, data mining, and performance analysis of computer networks.

Pilsung Kang is an assistant professor at School of Industrial Management Engineering, Korea University. He received his BS and Ph.D in Industrial Engineering from Seoul National University. His main research interest is in developing forecasting models using machine learning algorithms. He has also conducted research in application areas such as keystroke dynamics-based authentication, fault detection in manufacturing processes, and social network analysis. He has published a number of papers on related topics in leading journals such as *Pattern Recognition, Intelligent Data Analysis*, and *Neurocomputing*.