Machine Learning Applications on Box-Office Revenue Forecasting: The Taiwanese Film Market Case Study

Shih-Hao Lu[1], Hung-Jen Wang[1], and Anh Tu Nguyen[2](✉)

[1] Department of Business Administration, National Taiwan University of Science and Technology, Taipei 106, Taiwan
shlu@mail.ntust.edu.tw
[2] Department of Banking, Ho Chi Minh City University of Banking, Ho Chi Minh City 700000, Vietnam
tuna@hub.edu.vn

Abstract. The Random Forest algorithm (RFA) is used to predict the approximate final box-office revenue of a movie in the Taiwanese film market. The results show that the RFA has stable capabilities to predict the final box-office revenue of a movie during its theatrical period with an 80% overall accuracy. Two other machine learning algorithms, i.e., the Support Vector Machine and the Logistic Regression algorithms, are applied for comparison with the RFA. We find that the RFA still achieves the highest overall accuracy of prediction in our experiment. Additionally, we applied an unsupervised machine learning method to distinguish each group in the box office revenue categories in the classification problem. Also, the feature importance analysis indicates that word-of-mouth plays a vital role in theatrical revenue determination. Our findings imply several crucial suggestions for film distributors.

Keywords: Box-office revenue · random forest algorithm · support vector machine · logistic regression · self-organizing maps · Taiwanese film market

1 Introduction

The movie industry is one of the most dramatically growing industries internationally (Ghiassi, Lio, & Moon, 2015; Kim, Hong, & Kang, 2015). For instance, The Motion Picture Association of America (MPAA) reported that the box-office profits for all films had reached a high record of US$40.6 billion and a 5% increase in sales compared to 2016 (MPAA, 2017). Similarly, the Taiwanese film market also experienced significant growth with a total of 649 movies shown in cinemas, and the total box-office gross revenue totaled approximately US$0.34 billion with 43 million movie tickets sold in the 2017 calendar year. Nevertheless, not every film posted successful revenue (De Vany & Walls, 1999) since 30% of released movies break even and only 10% of movies make box-office profit (Hennig-Thurau, Houston, & Walsh, 2007). Therefore, from the view

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 384–402, 2024.
https://doi.org/10.1007/978-3-031-35763-3_49

of producers, distributors, and exhibitors, box-office forecasting is not only a difficult and challenging task but also an extremely important issue given that the results of these predictions directly determine their decision making (Delen, Sharda, & Kumar, 2007; Ghiassi et al., 2015; Hur, Kang, & Cho, 2016; Sharda & Delen, 2006).

A decade of research has now provided useful information on movie revenue forecasting using multiple machine learning algorithms. However, most research mainly focused on earnings for Hollywood movies or domestic earnings in the US (Brewer, Kelley, & Jozefowicz, 2009; Chintagunta, Gopinath, & Venkataraman, 2010; Dellarocas, Zhang, & Awad, 2007; Litman, 1983; Neelamegham & Chintagunta, 1999; Sawhney & Eliashberg, 1996), and other studies focused on target markets in the Chinese, Korean or Chilean film industry (Kim et al., 2015; Marshall, Dockendorff, & Ibáñez, 2013; L. Zhang, Luo, & Yang, 2009). Based on our best knowledge, total movie theater revenue and the development of forecasting models for the Taiwanese film industry using machine learning are still largely unknown and have not been investigated. Thus, we propose to employ the Random Forest algorithm (RFA) (Breiman, 2001), which has been examined to be an accurate approach in data classification (Lin, Wu, Lin, Wen, & Li, 2017; Sun, Zhong, Dong, Saeeda, & Zhang, 2017; Wu, Ye, Liu, & Ng, 2012; Ye, Wu, Huang, Ng, & Li, 2013), in our study.

This study differs from previous research based on the following factors. First, our prediction model is precise in up to 80% of all cases, which is one of the highest precision rates among box-office revenue predicting approaches to our knowledge. Second, we propose to directly collect, measure and use word-of-mouth (WOM) data to pursue feature importance, which is unaddressed in previous studies, and the final results indicate that WOM plays an active role in explaining our experiment. Finally, our study is one of the first attempts to predict theatrical revenue in Taiwan, one of the international developing markets for movies.

The paper is organized as follows. Section 2 briefly reviews the literature on forecasting box-office rentals. Section 3 provides the details of forecasting by RFA, including measurement of input variables and the RFA procedure in our experiment. The two sections that follow mainly provide empirical results along with a comparison of other approaches, including Support Vector Machine and Logistic Regression algorithm. The last two sections of the paper discuss some conclusions extracted from empirical evidence along with study limitations and further research suggestions.

2 Related Work

Some primary algorithm approaches, such as multiple regression models (Basuroy, Desai, & Talukdar, 2006; Brewer et al., 2009; De Vany & Walls, 1999; Duan, Gu, & Whinston, 2008a, 2008b; Elberse & Eliashberg, 2003; Eliashberg & Shugan, 1997; Litman, 1983; Litman & Kohl, 1989), Bayesian models (Ainslie, Drèze, & Zufryden, 2005; K. J. Lee & Chang, 2009; Neelamegham & Chintagunta, 1999), and machine learning algorithms (Delen & Sharda, 2010; Du, Xu, & Huang, 2014; Ghiassi et al., 2015; Hur et al., 2016; Kim et al., 2015; Sharda & Delen, 2006; L. Zhang et al., 2009; W. Zhang & Skiena, 2009), have been developed as reported in the literature on box-office revenue forecasting. Each approach has its own advantages. For example, the multiple regression models evaluate the importance of the variables, but variables must comply with the assumption of a normal or gamma distribution. Moreover, machine learning algorithms based on assessing nonlinear forecasting do not rely on these assumptions (Hur et al., 2016).

Regarding variable importance, Litman (1983) proposed the linear regression model with eight independent variables that are grouped into three main factors to determine a movies' theatrical success: creative sphere, scheduling and release pattern, and the marketing effort. The results show that production budget, distributor, time of release, Academy Award nominations and prizes, and critic reviews have positive effects on rentals of theatrical movies. Litman and Kohl (1989) added some new variables, including well-known ideas, country of origin, market forces, and advertising budget, to the three main factors in the model of (Litman, 1983) to examine the supply-side effect on adjusted rentals of theatrical movies. The results suggest that superstar power, distributor, positive reviews, summer season, and storyline drive movie revenues. Based on the three stages of the hierarchical Bayes model, Neelamegham and Chintagunta (1999) found that several factors, such as a number of screens showing a film, local distributors' impact on cinema earnings, and genre, were similar to separate geographic area. In the general forecasting results, the mean absolute percentage error of the prelaunch model of Neelamegham and Chintagunta (1999) is 36.6% lower than the proposed model of Sawhney and Eliashberg (1996) for the U.S. market. On the other hand, approaches based on machine learning have been developed recently, and the results show the overall accuracy of forecasting. Sharda and Delen (2006) developed the artificial neural network (ANN) model to predict a movie classified into one of nine categories from Flop to Blockbuster. The results show 36.9% classification accuracy for the exact (Bingo) hit rate. They also examined the overall accuracy as determined by traditional statistical classification methods, such as Logistic Regression (30.17%), Discriminant Analysis (29.25%), and Classification and Regression Tree (31.18%). Delen and Sharda (2010) analyzed four additional classification models in additional research to enhance the results of Sharda and Delen (2006). The overall accuracy results revealed 55.49% accuracy for the Support Vector Machine, 54.62% for Random Forest, 54.05% for Boosted Tree, and 56.07% for the Fusion (average). Recently, ongoing improvements in machine learning algorithms have led to many new and fascinating applications in box-office forecasting. Ghiassi et al. (2015) employed a dynamic artificial neural network model (DAN2) and argued that DAN2 has excellent performance with 94.1% accuracy. In addition, Ghiassi et al. (2015) eliminated variables, such as competition, star value, and special effects, from DAN2 and replaced these variables with production budget, pre-released advertising expenditures, runtimes, and seasonality. As a consequence, DAN2 exhibited better performance than ANN, as assessed by Delen and Sharda (2010). W. Wang, Xiu, Yang, and Liu (2018) applied a deep belief network (DBN) model to predict box office revenue in China. The experimental results revealed that the DBN had the lowest mean absolute error and root mean square error, comparing the traditional BRP model and the back-propagation neural network. In another research for the Chinese movie market, Liao, Peng, Shi, Shi, and Yu (2020) showed that the applied stacking fusion model performed Bingo and 1-Away accuracy at 69% and 86%, respectively.

3 Forecasting with the Random Forest Algorithm

A two-phase study was designed to validate whether the prediction models applied in this research can evaluate variable importance robustly within the dataset and authenticate the accuracy of the proposed forecasting framework. In this study, one algorithm is employed to build the prediction model, and two additional different algorithms are utilized to compare the prediction model's performance as shown in Fig. 1.

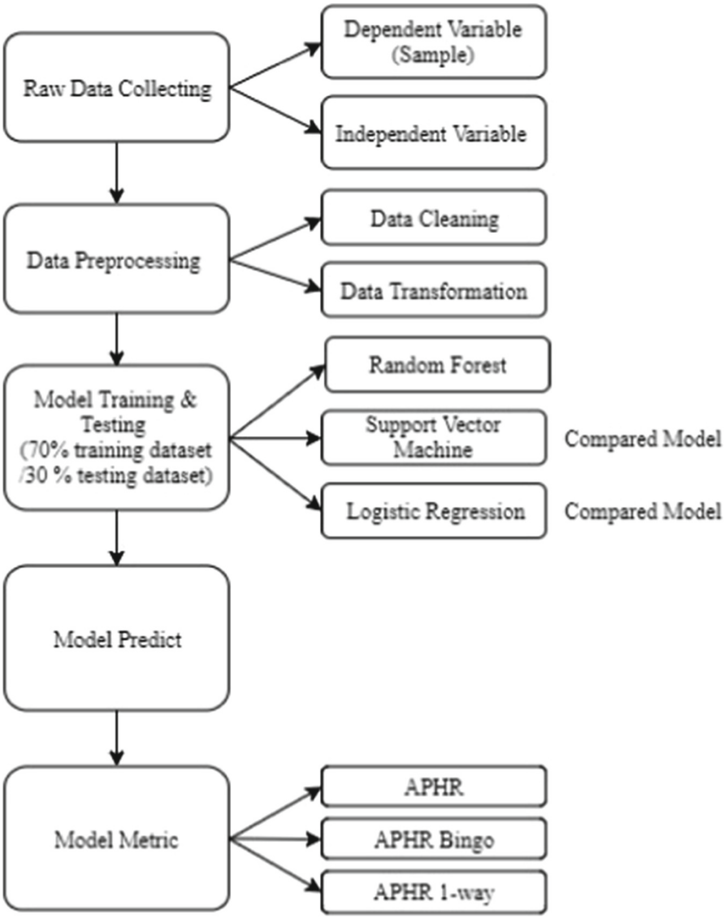


Fig. 1. Illustration of the forecasting process

3.1 Raw Data Collecting

Dependent Variable. To validate variable importance and achieve the highest accuracy of the model, the dataset must consist of the complete calendar year information. We gathered 498 movies released between January 2017 and May 2018 from the weekly report conducted by National Taiwan Film Institute as sample data. The dependent variable is the box-office revenue which is ranged from 3,150 New Taiwan Dollars (NT$) to 641 million NT$. We therefore divide box-office revenue into several categories to figure out the relationship between dependent and independent variables. However, unlike previous studies (Delen and Sharda (2010), Ghiassi et al. (2015), and Sharda and Delen (2006)) which are based on expert opinions to classify groups, we applied the self-organizing maps neural network (SOM) clustering method proposed by Kohonen (1982) to deal with group classification. SOM clustering is the unsupervised classification algorithm which is widely applied for dealing with several issues in engineering and data analysis to diagnose label of items (Markonis & Strnad, 2020; Schmidt, Rey, & Skupin, 2011). SOM clustering includes input vector and one layer of network topology which consists output neurons. The input vectors $i = [i_1, i_2, \ldots i_n]$ are fed to the system by linear transfer function and each input node connects to the output neurons by weighted average $w_i = [w_{i1}, w_{i2} \ldots w_{in}]$. The outcome is determined by finding a neuron which is its best matching unit (Nanda, Sahoo, & Chatterjee, 2017). The SOM has also been identified by some parameters, i.e., neighborhood area, neighborhood coefficient, and neighborhood shrinking. More specifically, the first one represents for the 2-dimensional network topology of output nodes, the second one implies for the parameter controls the interaction of output nodes inside the neighborhood area, and the third one means the neighborhood radius decreases after each iteration to figure out the best matching unit. The SOM clustering was coded by the Matlab R2019b software. The detailed procedure with all parameters set up was presented in Appendix. The results show that box-office earning in our sample should be classified into six categories as shown in Table 1.

Table 1. Output Variables Classified Thresholds

Class No	Range (in 10 thousands NT$)	Numbers of samples
A	< 500 (Flop)	352
B	500–999	37
C	1,000–1,999	26
D	2,000–3,999	28
E	4,000–9,999	22
F	> 10,000 (Blockbuster)	33

Independent Variables. Seven different types of independent variables were used, including six variables categorized as internal variable extraction and one variable (WOM) classified as external variable extraction. With regard to the internal variables, we referred to previous studies in the literature (Ghiassi et al., 2015; Hur et al., 2016; Kim et al., 2015; Litman, 1983; Litman & Kohl, 1989; Sharda & Delen, 2006) and collect data from sources as follows.

MPAA rating: In the U.S., before a film is officially released on the screen, it is assigned a rating of G, PG, PG-13, R, or NC-17 based on suitability for audiences with regard to violence and sexual problems. Therefore, the input variables from The MPAA rating system is one of the most widely utilized variables since it is an awareness system and its ratings and their definitions have remained relatively static over the past few decades (Ghiassi et al., 2015). Moreover, each particular rating decision might hold additional predictive power for box-office revenue forecasting (Ghiassi et al., 2015) because these ratings emit signals regarding film content that moviegoers find informative for personal decision making (Prag & Casavant, 1994). However, correlation results between MPAA and box-office success were divergent. Some research indicated that MPAA has a partial influence (Dellarocas et al., 2007) or even no significant influence on box-office revenue (Litman, 1983; Litman & Kohl, 1989). Meanwhile, W. Zhang and Skiena (2009) report the correlation between a movie's rating and its gross revenue. Prag and Casavant (1994) indicate that the movie ratings (G, PG, PG13, and R) have significant positive impacts on movie rental, and the MPAA ratings easily disclose movie content for films without a large budget for advertising. In this study, we use five binary variables based on Taiwan's movie rating system, which uses categories of G, P6, PG12, PG15, and R as substitutes for the MPAA ratings. Accordingly, appropriate moviegoers are classified by their ages; for example, the P6-class prohibits children under 6 and requires accompanying parents or adult guardians for children under 12.

Genre: Previous research commonly used movie genre as an important input variable for the theatrical success forecasting models; however, it is difficult to determine the impact of a genre on revenue because a film can be classified into multiple genres (Ghiassi et al., 2015). As a consequence, these models rarely found a significant relationship between genre and movie success or briefly classified a film based on specific genres (Litman, 1983; Litman & Kohl, 1989; Sharda & Delen, 2006; L. Zhang et al., 2009). Nevertheless, some research provided more information and pointed out that science fiction, horror, and comedy were the main movie categories associated with movie theater revenue (Litman, 1983; Liu, 2006). Furthermore, the genre of a film is an important attribute because it can determine the prospective audience demographics combined with the film rating or release timing. For example, the prospective audience for a G-rated family movie is different compared to an R-rated thriller movie, or different genres released around continuing holidays or on a particularly significant day can result in different gross revenue (Ghiassi et al., 2015). In this study, we used 19 binary-independent variables (Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, History, Horror, Musical, Mystery, Romance, Science-Fiction, Sport, Thriller, War) that follow the convention reported by Sharda and Delen (2006) and gathered these variables from the IMDb website.

Distributor: To provide a more multifaceted perspective, the motion picture industry is required to make appropriate managerial decisions when distributing a film to theaters to maximize revenue (Hur et al., 2016). Moreover, the greatest distributors are more likely to announce a signal of good quality and increase film receipt (Gong, Van der Stede, & Mark Young, 2011). Litman (1983) supposed that major distributors have some advantages to produce and distribute a film to audiences given considerable financing and good connection to exhibitors' networks. As a consequence, Litman (1983) found empirical evidence that supports the hypothesis that major distributors enhance rentals of theatrical movies. In addition, Basuroy et al. (2006) reconfirmed the hypothesis of Elberse and Eliashberg (2003) that major distributors indirectly affect revenue by increasing film screens during opening week. Generally, it can be assumed that distributor power is one of the factors that is more likely to drive a film's success. In our study, this variable is divided into three binary variables: high influence distributors, medium influence distributors, and others.

Sequel: In previous studies, empirical evidence showed that sequel movies correlate with the financial success of a movie (Dhar, Sun, & Weinberg, 2012; Ghiassi et al., 2015; Moon, Bergey, & Iacobucci, 2010; Sharda & Delen, 2006) even though a sequel has low quality and fewer stars (Ravid, 1999). Additionally, K. Lee, Park, Kim, and Choi (2018) noted that movie producers often produce sequel movies to reduce risk and uncertainty. In our study, the empirical model included a binary variable to identify whether a film is a sequel that is also widely used in further research (Ravid, 1999; Sharda & Delen, 2006).

Seasonality: An appropriate schedule for a theatrical movie might be crucial because it can likely influence the financial success of a film. One reason is that more moviegoers prefer to choose to watch a movie in their leisure time and a common film would gain great financial success during an important season, such as weekends (Duan et al., 2008a), summer months (Brewer et al., 2009; Litman, 1983), spring festival (L. Zhang et al., 2009), or the Christmas holiday (Gong et al., 2011; W. Zhang & Skiena, 2009). Commonly, film studios will schedule a theatrical movie to maximize the box-office revenue during long holidays for celebration. In this research, seasonality is measured by a binary factor that is coded 1 if a film is released on a long holiday (3 days or more) or 0 if not.

Nationality: Some earlier empirical studies indicated that a film origin could impact its movie theater success. For example, F. Wang, Zhang, Li, and Zhu (2010) discovered that movies produced in China have a significantly positive effect on aggregative box-office revenue. In contrast, L. Zhang et al. (2009) found that international movies can be more profitable than Chinese movies that are screened, produced, and filmed in China, and W. Zhang and Skiena (2009) showed that movies originating from the USA exhibit a clearly significant correlation with movie revenue. In Taiwan, the top ten profitable movies in 2017 were all from Hollywood. These facts suggest that a movie's financial success is more likely correlated with the place where the movie originates, and this correlation is especially significant for Hollywood movies in the Taiwanese market. In our study, we divided movie origin into seven binary variables: Hollywood, Chinese, Japanese, Korean, Thai, Bollywood, and others.

In addition to input variable features extracted from movies, the inclusion of a set of word-of-mouth (WOM) external variables in the forecasting model is imperative because it can positively influence the accuracy of the model (Asur & Huberman, 2010; Duan et al., 2008a; Liu, 2006). Liu (2006) supposed that WOM determines movie revenue in two phases. First, WOM volume increases filmgoers' awareness. Second, WOM valence affects consumers' attitudes about the films and their decisions making. However, the results in the literature are quite divergent. Some earlier studies advocate that WOM exhibits a positive contribution to the film industry (Baek, Oh, Yang, & Ahn, 2017; Du et al., 2014; Duan et al., 2008a, 2008b; Elberse & Eliashberg, 2003), whereas other studies showed a partial influence on revenue (Basuroy, Chatterjee, & Ravid, 2003; Chintagunta et al., 2010; Duan et al., 2008b; Eliashberg & Shugan, 1997; Liu, 2006). For example, Elberse and Eliashberg (2003) argue that WOM is a crucial predictor of revenue and screens in subsequent weeks. Liu (2006) found that the volume of WOM is the most significant effect on theatrical rentals, whereas the valence of WOM is not. Similarly, Duan et al. (2008b) demonstrated that WOM valence indirectly increases box-office revenue by generating a higher volume of WOM rather than directly influencing revenue. Furthermore, scholars have recently focused on the effects of electronic WOM in the era of the Internet revolution given the speed of WOM transmission (Duan et al., 2008b). For example, Duan et al. (2008a, 2008b) tracked the data from three different social network service platforms, including Yahoo! Movies, Variety.com, and BoxOfficeMojo.com, and summarized the daily and the cumulative number of posts for each movie. Asur and Huberman (2010) collected Twitter posts per hour that mentioned a specific movie as the input data to assess the volume level. In this study, to avoid inaccurate predictions due to the omission of WOM effects, we divided the WOM variable into two components: WOM volume with three indicators (the rating of a movie, the quantity of voters for a movie, and the total views of a movie trailer) and WOM valance with two indicators (the total number of likes/dislikes for a movie trailer). WOM data are collected from IMDb, Yahoo! Movies TW, and YouTube (Table 2).

Table 2. Summary of Independent Variables Extracted from the Movie Aspect

Independent Variables	No. of values	Description	Data source	Independent Variables	Classification	Description	Data source
Internal Variables				External Variables			
Movie Rating	5	G, P6, PG12, PG15, R	Taiwan BAMID	The rating of a movie	Volume	Positive Integer	IMDb
Distributor	3	High, Medium influence distributors, others	NTFI	The no. of voters for a movie	Volume	Positive Integer	IMDb
Nationality	7	Hollywood, Chinese, Japanese, Korean, Thai, Bollywood, others	NTFI	The rating of a movie	Volume	Positive Integer	Yahoo! Movies TW
The Official Release Schedule	2	Continuous holiday (3 days above), No	NTFI	The no. of voters for a movie	Volume	Positive Integer	Yahoo! Movies TW
Sequel	2	Yes, No	NTFI	The total views of a movie trailer	Volume	Positive Integer	YouTube

(*continued*)

Table 2. (*continued*)

Independent Variables	No. of values	Description	Data source	Independent Variables	Classification	Description	Data source
Genre	19	Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Horror, Mystery, Romance, Sci-Fi, Sport, Thriller, History, Musical, War	NTFI	The total number of likes for a movie trailer	Valence	Positive Integer	YouTube
				The total number of dislikes for a movie trailer	Valence	Positive Integer	YouTube

3.2 Data Preprocessing

To enhance the accuracy of the model, two major data preprocessing points were applied: data cleaning and data transformation. The former relates to filling in missing values, dropping outliers, and resolving inconsistencies in the data. The latter pertains to adjusting different dimensions and increasing the accuracy of the forecasting models. In particular, data cleaning is used to remove movies merely released to the Taiwanese market because it is difficult to link and extract corresponding voting and rating data at IMDb. In addition, data cleaning is used to add the appropriate values, which are derived from a sample with similar features within the research dataset. For example, the movie '*Jump! Man*' lacks the value of the number of voters on IMDb; however, the movie also belongs to the Taiwanese documentary movie dataset.

Data transformation, on the other hand, is a process that normalizes raw data with different meanings, dimensions, units, or scales to properly format data for the forecasting model. In our study, dummy variables are numbered via one-hot encoding, and the numerical variables are normalized to format the features within the raw dataset.

3.3 Model Training and Testing

In this study, we used the Average Percent Hit Rate (APHR), which is the most intuitive indicator to measure the discrimination for the predictive accuracy of a classification problem. We also applied two performance measures as the prediction results of the model: the average percent hit rate of exactly classifying a movie's success (Bingo) and

1-Away. The APHR indicator, the Bingo, and the 1-Away were introduced by Sharda and Delen (2006). More specifically, the APHR measures the ratio of correct classifications to the total number of movies in the sample. The bigger values of APHR, the better the classification performance. The Bingo is applied to precisely classify a movie into one of six categories based on revenue thresholds, whereas the 1-Away is allowed for two subsequent categories, as shown in Table 1. For example, if a movie revenue is predicted to group A (revenue is less than 5 million NT$), but the actual revenue of the movie is B (revenue is between 5 to 10 million NT$), the precision for the Bingo is incorrect meanwhile for 1-Away is correct. In other words, the Bingo shows the exact prediction while the 1-Away allows predicted values for a broader range that may reflect real scenarios (Delen & Sharda, 2010; Ghiassi et al., 2015; Sharda & Delen, 2006).

For sample model training and dataset testing, previous studies showed that using a single experiment or a single method was inappropriate, and the subsequent use of k-fold cross-validation is ideally appropriate (Sharda & Delen, 2006). Nevertheless, K. Lee et al. (2018) demonstrated that this approach could deteriorate if the volume of data is small. If the dataset is small, repeated random subsampling validation is more suitable than k-fold cross-validation in our samples. As such, we repeat the validation process ten times using 70% of samples as training data and the remaining 30% of samples as testing data.

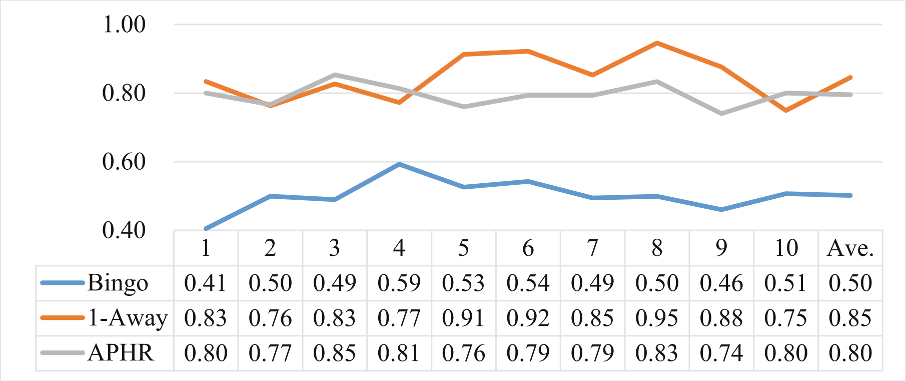
4 Results

The confusion matrix presented in Table 3 is an example of one out of the classification results of testing data subject to ten times of iteration. The columns represent the actual classes, whereas the rows represent the predicted classes in the confusion matrix. The correct classification of the samples for that class is presented in the intersection cells of the same classes.

Table 3. A Confusion Matrix Example of One Result from Ten-Times Iterated Classification of Testing Data

		Actual						Avg.
		A	B	C	D	E	F	
Predicted	A	109	1	1	0	0	1	
	B	7	3	0	1	0	0	
	C	5	2	0	1	0	0	
	D	1	0	2	0	2	2	
	E	1	0	0	1	2	2	
	F	0	0	0	0	0	6	
	Bingo	0.89	0.50	0.00	0.00	0.50	0.55	0.41
	1-Away	0.94	1.00	0.67	0.67	1.00	0.73	0.83
	Average Percent Hit Rate (APHR)							0.80

Table 3 also reveals the prediction accuracy of each class individually and overall prediction accuracy in the lower column. For instance, in this iteration of the prediction process, 109/150 samples were accurately predicted to be class A compared with real results, while others represent the misclassifications. Thus, the highest hit rates of Bingo and 1-Away for class A are 0.89 and 0.94, respectively and are the highest exact proportions among movie types. In addition, the overall APHR is 0.80, which indicates that the prediction model is able to classify 80% of samples correctly into their classes within this experiment. Furthermore, the aggregated ten-fold iterated classification results in Fig. 2 show that the average overall accuracy of APHR is 0.80, whereas the prediction accuracy of Bingo and 1-Away is 0.50 and 0.85, respectively. The obtained results are better than those in Sharda and Delen (2006), which were 37% for APHR, 37% for Bingo, and 76% for 1-Away. Although the accuracy of the Bingo metric is relatively low (50%) in our experiments, the 1-Away metric reaches 85%, indicating a practical approach for practical application, as previously mentioned.



	1	2	3	4	5	6	7	8	9	10	Ave.
Bingo	0.41	0.50	0.49	0.59	0.53	0.54	0.49	0.50	0.46	0.51	0.50
1-Away	0.83	0.76	0.83	0.77	0.91	0.92	0.85	0.95	0.88	0.75	0.85
APHR	0.80	0.77	0.85	0.81	0.76	0.79	0.79	0.83	0.74	0.80	0.80

Fig. 2. An aggregated ten-times iterated classification result

To improve the accuracy of the prediction models, we perform feature importance analysis to determine the independent variable(s) that mostly affect dependent variables in the proposed models. The results are summarized in Fig. 3. Here, the x-axis represents the input variables, and the y-axis represents the importance of the input variables. Altogether, the majority of external variables (WOM) have significant contributions to the prediction of a movie's financial success. For instance, all of these variables have important explanations for box-office revenue compared with internal variables. Additionally, most of the explanatory power is derived from the volume of WOM, which is consistent with the findings of Duan et al. (2008a, 2008b), Liu (2006). Taiwanese filmgoers mostly use information from Yahoo! Movie TW to learn about a movie.

Figure 3 also presents a point of comparison of overall APHR for RFA without internal or external variables. As the line graph suggests, the dataset without the internal variables for RFA can also result in the same overall prediction accuracy (0.8). In contrast, the dataset without the internal variables has decreased by 8%. This finding reconfirms the important contribution of WOM in our forecasting model.

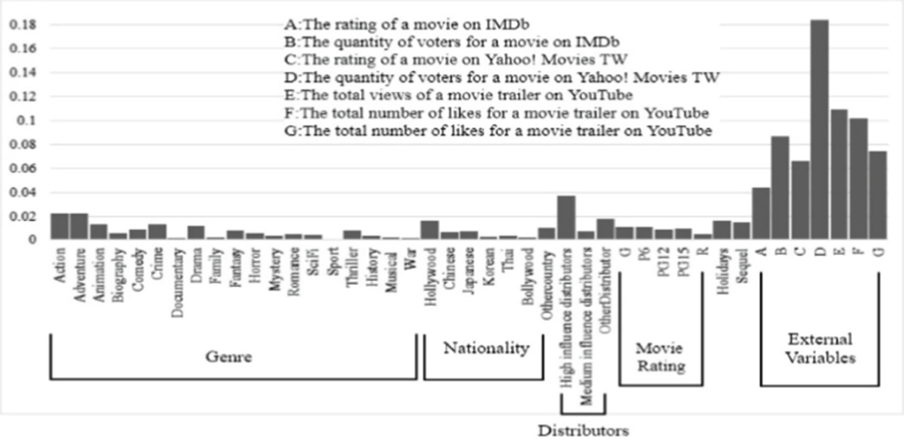
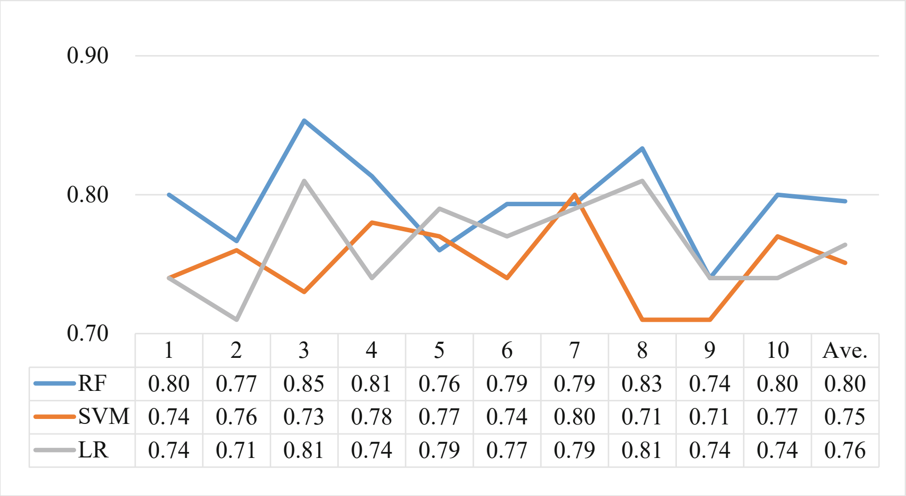


Fig. 3. Feature importance analysis

5 Comparison to Other Models

Two additional machine learning algorithms, Support Vector Machine algorithm (SVM) and Logistic Regression algorithm (LR), were applied to validate the performance of the prediction of the RF model. SVM is expected to identify the maximum margin hyperplanes that optimally classify the categories in the training data (Delen & Sharda, 2010; K. Lee et al., 2018), while LR is used to predict binary or multiclass dependent variables (K. Lee et al., 2018; Sharda & Delen, 2006). Using the same training and testing dataset with the same cross-validation method for RFA, SVM, and LR, Fig. 4 presents



	1	2	3	4	5	6	7	8	9	10	Ave.
RF	0.80	0.77	0.85	0.81	0.76	0.79	0.79	0.83	0.74	0.80	0.80
SVM	0.74	0.76	0.73	0.78	0.77	0.74	0.80	0.71	0.71	0.77	0.75
LR	0.74	0.71	0.81	0.74	0.79	0.77	0.79	0.81	0.74	0.74	0.76

Fig. 4. An aggregated ten-times iterated classification result (APHR) for random forest, support vector machine, and logistic regression algorithm

the results for the ten-fold iteration for each approach. As the line graph shows, the RF has a higher average overall accuracy with an APHR of 0.80 than the SVM algorithm (0.75) and LR algorithm (0.76). Therefore, the RF better performs classification tasks than SVM or LR in this research. We note that the results obtained in this study are higher than those in Sharda and Delen (2006) and are the same accuracy as those in Liao et al. (2020).

6 Conclusion and Discussion

Some findings extracted from this study could be useful for distributors in Taiwan to determine the financial success of a movie. First, the results show that the RF has stable capabilities to predict the final box-office revenue of a movie during its theatrical period within an 80% rate of accuracy. Additionally, a comparison result of this validation demonstrates that RF achieves the highest average overall accuracy (APHR) compared to others (SVM and LR) in this research. This contribution provides a detailed framework of RF for future researchers or practical distributors in Taiwan in the field of forecasting box-office revenue. Second, in our proposed model, WOM plays a crucial role in cinema success, which is consistent with the findings of Baek et al. (2017). Given that WOM has extraordinary transmission speed through the Internet (Duan et al., 2008b) and that an appropriate marketing strategy before or after a movie's release is vital (Ghiassi et al., 2015), distributors could deploy advertising campaigns on social networks channels (IMDb, Yahoo! TV Taiwan) before a film's release to increase WOM volume or increase interaction at movie review forums (YouTube) to boost WOM valence during a film's performance at the box-office. In addition, although the internal variables have inconclusive contributions to box-office forecasting, these variables still provide different suggestions for the decision-makers to help a movie be successful. Our collected data revealed that successful movies likely have similar features. For example, if a movie is a Hollywood action or adventure movie, such as "*The Avengers*", and released on an important holiday, such as Chinese New Year or during Winter or Summer break, this movie is more likely to achieve over NT$100 million at the box-office in the Taiwanese market.

7 Limitations and Future Research

We acknowledge three primary limitations involved with the use of machine learning algorithms to solve the forecasting problem within this study. The first limitation is the lack of sufficient data for the forecasting model. It is difficult to collect the exact number of total box-office sales from movie contributors, and this research relies upon data gathered from the National Taiwan Film Institute. Thus, data are potentially not accurate, and the proposed model might not reflect the real world. In addition, the movie market in Taiwan is narrow and classes B to E only represent 30% of samples compared to the A class, which represents 70% of samples. Consequently, unbalanced data might result in a reduction in the model's performance and might not produce reasonable results for these classes within this experiment. Second, concerning WOM variable reliability, the rating and the volume of voting specifically represent the audience's perspective. It is difficult to measure whether the ratings are real or fake. The results therefore could be biased. This issue is also for any forecasting model in future research. Finally, the official release schedule in Taiwan for some movies is occasionally ahead of the standard schedule in Hollywood, leading to insufficient information related to movies that can be used for the prediction models, such as the IMDb rating. Although the missing values can be completed using a variety of data preprocessing techniques, the results could be inaccurately reflected.

With regard to future research, some work is needed to improve our results. From the perspective of the variables used in the model, although our results reach 80% accuracy for predicting real cases, researchers could add different features based on the movie's aspects, i.e., the star value, the number of screens, or the special effects, to improve the final prediction results. In addition, other research suggests that the researcher can include a variety of WOM variables for comparison, such as the data compiled from Google trends (Panaligan & Chen, 2013) or other popular social media platforms. From the utilized approaches based on machine learning algorithms, we suggest that future research can explore different techniques to address the prediction problem for the movie domain within the Taiwanese market, such as a backpropagation algorithm.

Appendix

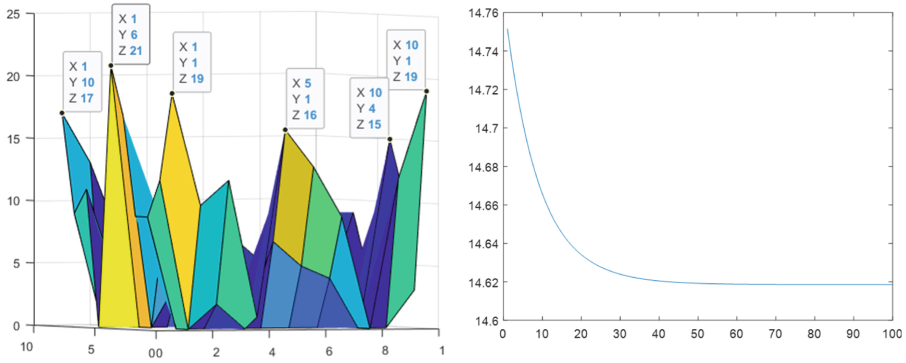
Self-Organizing Maps Algorithm Pseudocode for Box Office Revenue in Taiwan

```
Input Training: network parameters
Network parameters setting
initialize: 2-dimensional topology, connecting weight matrix randomly
iteration = 0
input node = 43
number of sample = N
neighborhood coefficient ($R = 10\sqrt{2}, 8\sqrt{2}$)
learning rate (n = 0.01, 0.1, 0.9)
for iteration: 1:100
  for N = 1:N
      generate zeros 2-dimensional topology
    for j = 1: j (x-axis topology)
      for k = 1: k (y-axis topology)
        for i = 1: i (input node)
          Calculating the net value between input and output topology
       Selecting the minimum node as best matching unit
        end
      end
    end
    Calculating the output vector Y for each output layer
    for j = 1: j (x-axis topology)
      for k = 1: k (y-axis topology)
        for i = 1: i (input node)
        $\Delta w=n*(x(N,i)-w(i,j,k))*\exp(-\mathrm{sqrt}((j-j_{bmu})^2+(k-k_{bmu})^2)/R)$
        $w(i,j,k)=w(i,j,k)+\Delta w(i,j,k)$
        end
      end
    end
  end
    for j = 1: j (x-axis topology)
      for k = 1: k (y-axis topology)
        for i = 1: i (input node)
        Calculating error for each input node
        Minimizing the sum of error
        end
      end
```

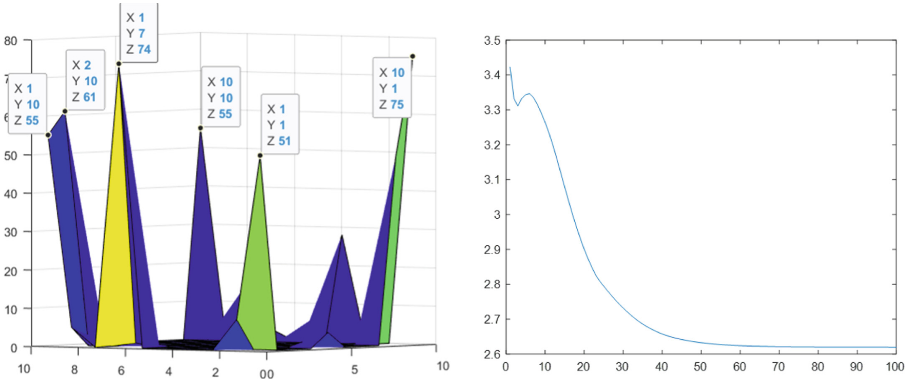
```
        end
      Shrinking learning rate
      Shrinking neighborhood radius
  end
  Input Recalling
  Network parameters setting as training phase
  The same procedure as training phase
```

Self-Organizing Maps Clustering Results and Illustration of Training Error

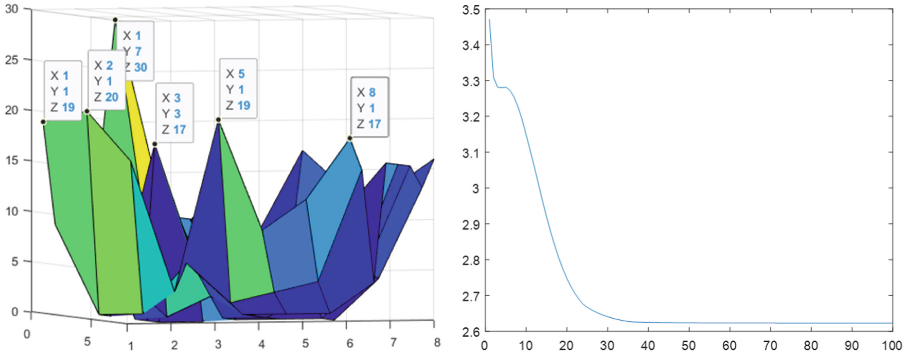
Case 1: Topology: 10*10, neighborhood radius: 10 $\sqrt{2}$, shrink learning rate: 0.9; iteration: 100.



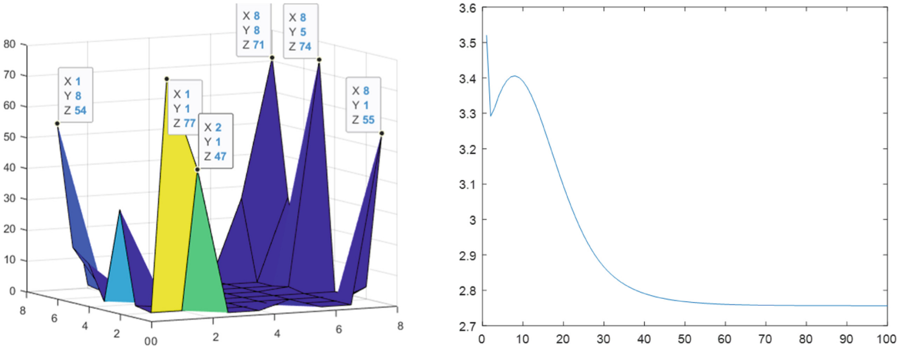
Case 2: Topology: 10*10, neighborhood radius: 10 $\sqrt{2}$, shrink learning rate: 0.01; iteration: 100.



Case 3: Topology: 8*8, neighborhood radius: 8 $\sqrt{2}$, shrink learning rate: 0.9; iteration: 100.



Case 4: Topology: 8*8, neighborhood radius: 8 $\sqrt{2}$, shrink learning rate: 0.01; iteration: 100.



References

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

Asur, S., Huberman, B.A.: 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 (2010)

Baek, H., Oh, S., Yang, H.-D., Ahn, J.: Electronic word-of-mouth, box office revenue and social media. Electron. Commer. Res. Appl. **22**, 13–23 (2017)

Basuroy, S., Chatterjee, S., Ravid, S.A.: How critical are critical reviews? The box office effects of film critics, star power, and budgets. J. Mark. **67**(4), 103–117 (2003)

Basuroy, S., Desai, K.K., Talukdar, D.: An empirical investigation of signaling in the motion picture industry. J. Mark. Res. **43**(2), 287–295 (2006)

Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

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

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

De Vany, A., Walls, W.D.: Uncertainty in the movie industry: does star power reduce the terror of the box office? J. Cult. Econ. **23**(4), 285–318 (1999)

Delen, D., Sharda, R.: Predicting the financial success of Hollywood movies using an information fusion approach. Indus. Eng. J. **21**(1), 30–37 (2010)

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

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

Dhar, T., Sun, G., Weinberg, C.B.: The long-term box office performance of sequel movies. Mark. Lett. **23**(1), 13–29 (2012)

Du, J., Xu, H., Huang, X.: Box office prediction based on microblog. Expert Syst. Appl. **41**(4), 1680–1689 (2014)

Duan, W., Gu, B., Whinston, A.B.: Do online reviews matter?—an empirical investigation of panel data. Decis. Support Syst. **45**(4), 1007–1016 (2008)

Duan, W., Gu, B., Whinston, A.B.: The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry. J. Retail. **84**(2), 233–242 (2008)

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

Eliashberg, J., Shugan, S.M.: Film critics: influencers or predictors? J. Mark. **61**(2), 68–78 (1997)

Ghiassi, M., Lio, D., Moon, B.: Pre-production forecasting of movie revenues with a dynamic artificial neural network. Expert Syst. Appl. **42**(6), 3176–3193 (2015)

Gong, J.J., Van der Stede, W.A., Mark Young, S.: Real options in the motion picture industry: evidence from film marketing and sequels. Contemp. Account. Res. **28**(5), 1438–1466 (2011)

Hennig-Thurau, T., Houston, M.B., Walsh, G.: Determinants of motion picture box office and profitability: an interrelationship approach. RMS **1**(1), 65–92 (2007)

Hur, M., Kang, P., Cho, S.: Box-office forecasting based on sentiments of movie reviews and Independent subspace method. Inf. Sci. **372**, 608–624 (2016)

Kim, T., Hong, J., Kang, P.: Box office forecasting using machine learning algorithms based on SNS data. Int. J. Forecast. **31**(2), 364–390 (2015)

Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**(1), 59–69 (1982)

Lee, K., Park, J., Kim, I., Choi, Y.: Predicting movie success with machine learning techniques: ways to improve accuracy. Inf. Syst. Front. **20**(3), 577–588 (2018)

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

Liao, Y., Peng, Y., Shi, S., Shi, V., Yu, X.: Early box office prediction in China's film market based on a stacking fusion model. Ann. Oper. Res. **308**, 1–18 (2020)

Lin, W., Wu, Z., Lin, L., Wen, A., Li, J.: An ensemble random forest algorithm for insurance big data analysis. IEEE Access **5**, 16568–16575 (2017)

Litman, B.R.: Predicting success of theatrical movies: an empirical study. J. Pop. Cult. **16**(4), 159–175 (1983)

Litman, B.R., Kohl, L.S.: Predicting financial success of motion pictures: the'80s experience. J. Media Econ. **2**(2), 35–50 (1989)

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

Markonis, Y., Strnad, F.: Representation of European hydroclimatic patterns with self-organizing maps. Holocene **30**(8), 1155–1162 (2020)

Marshall, P., Dockendorff, M., Ibáñez, S.: A forecasting system for movie attendance. J. Bus. Res. **66**(10), 1800–1806 (2013)

Moon, S., Bergey, P.K., Iacobucci, D.: Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. J. Mark. **74**(1), 108–121 (2010)

MPAA. A comprehensive analysis and survey of the theatrical and home entertainment market environment (Theme) for 2017 - THEME REPORT (2017). https://www.mpaa.org/wp-content/uploads/2018/04/MPAA-THEME-Report-2017_Final.pdf

Nanda, T., Sahoo, B., Chatterjee, C.: Enhancing the applicability of Kohonen self-organizing map (KSOM) estimator for gap-filling in hydrometeorological timeseries data. J. Hydrol. **549**, 133–147 (2017)

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

Panaligan, R., Chen, A.: Quantifying movie magic with google search. Google Whitepaper—Industry Perspectives+ User Insights (2013)

Prag, J., Casavant, J.: An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. J. Cult. Econ. **18**(3), 217–235 (1994)

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

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

Schmidt, C.R., Rey, S.J., Skupin, A.: Effects of irregular topology in spherical self-organizing maps. Int. Reg. Sci. Rev. **34**(2), 215–229 (2011)

Sharda, R., Delen, D.: Predicting box-office success of motion pictures with neural networks. Expert Syst. Appl. **30**(2), 243–254 (2006)

Sun, J., Zhong, G., Dong, J., Saeeda, H., Zhang, Q.: Cooperative profit random forests with application in ocean front recognition. IEEE Access **5**, 1398–1408 (2017)

Wang, F., Zhang, Y., Li, X., Zhu, H.: Why do moviegoers go to the theater? The role of prerelease media publicity and online word of mouth in driving movie going behavior. J. Interact. Advert. **11**(1), 50–62 (2010)

Wang, W., Xiu, J., Yang, Z., Liu, C.: A deep learning model for predicting movie box office based on deep belief network. In: Tan, Y., Shi, Y., Tang, Q. (eds.) ICSI 2018. LNCS, vol. 10942, pp. 530–541. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93818-9_51

Wu, Q., Ye, Y., Liu, Y., Ng, M.K.: SNP selection and classification of genome-wide SNP data using stratified sampling random forests. IEEE Trans. Nanobiosci. **11**(3), 216–227 (2012)

Ye, Y., Wu, Q., Huang, J.Z., Ng, M.K., Li, X.: Stratified sampling for feature subspace selection in random forests for high dimensional data. Pattern Recogn. **46**(3), 769–787 (2013)

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

Zhang, W., Skiena, S.: Improving movie gross prediction through news analysis. In: 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (2009)