Heliyon 6 (2020) e04260

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Comparing performance of ensemble methods in predicting movie box office revenue

Sangjae Lee [a,*], Bikash KC [a], Joon Yeon Choeh [b]

[a] College of Business Administration, Sejong University, Seoul 05006, South Korea
[b] Department of Software, Sejong University, Seoul 05006, South Korea

ARTICLE INFO

Keywords:
Movie box office revenue
Ensemble methods
Prediction of box office revenue
Decision trees
Data analysis
Data analytics
Big data
Management
Business management

ABSTRACT

While many business intelligence methods have been applied to predict movie box office revenue, the studies using an ensemble approach to predict box office revenue are almost nonexistent. In this study, we propose decision trees, k-nearest-neighbors (k-NN), and linear regression using ensemble methods and the prediction performance of decision trees based on random forests, bagging and boosting are compared with that of k-NN and linear regression based on bagging and boosting using the sample of 1439 movies. The results indicate that ensemble methods based on decision trees (random forests, bagging, boosting) outperform ensemble methods based on k-NN (bagging, boosting) in predicting box office at week 1, 2, 3 after release. Decision trees using ensemble methods provide better prediction performance than ensemble methods based on linear regression analysis in the box office at week 1 after release. This is explained by the results that after comparing the prediction performance between ensemble methods and non-ensemble methods. For decision tree methods, unlike the other methods, the prediction performance of ensemble methods is greater than that of non-ensemble methods. This shows that decision trees using ensemble methods provide better application effectiveness of ensemble methods than k-NN and linear regression analysis.

1. Introduction

The prediction of movie box office revenue after release has always been a challenging problem in a movie industry. The movie industry is growing day by day in a rapid way locally and globally. Hundreds and thousands of movies are released every month and a year (Kim et al., 2015). The correct prediction of box office revenues is important for the development of the movie industry and to lessen the market risk.

The success and failure of the movie depend on movie related variables such as timing of a movie release (i.e. movies released in high or low season) whether it is a sequel or not. eWOM (online word-of-mouth) have been used to predict box office revenue and these are provided through many forms such as online reviews, discussion boards, video sites, blogs, micro blogs, social networks etc(Baek et al., 2017; Qin, 2001; Oh et al., 2017). eWOM is becoming available in a large amount in this age of big data; and we are increasingly surrounded by enormous data available about the movies over the internet for which it is better to handle with business intelligence (BI) methods to process, manage and utilize these large data sets properly (Guille and Hacid, 2012). eWOM include positive reviews than negative reviews (helpfulness), number of reviews in the early stage of release or after the release, total helpful votes of the reviewers, an advertisement of a movie prior to release etc (Leenders and Eliashberg, 2011). In the previous studies, using eWOM and movie related variables, many researchers have tried to build up predicting box office revenue mainly using statistical regression algorithms such as multiple linear regression (Asur and Huberman, 2010) or machine learning algorithms, multi-layer perceptron neural network model (Ru et al., 2018; Wang et al., 2020), Bayesian belief network and backpropagation (BP) neural network which have robust performance and eradicates the limitations of the regression method with better prediction accuracy and are composed to predict the box office performance (Lee and Chang, 2009; Sharda and Delen, 2006; Zhang et al., 2009).

The study has the following motivations. First, as there exist a lack of studies regarding how ensemble methods can improve box office predict, we intend to fill this void by proposing the ensemble methods of decision trees, k-nearest-neighbors (k-NN), and linear regression to achieve improved predictive accuracy of box office earnings for one, two, three weeks after release. While neural networks have been utilized to predict box office effectively to accomplish satisfactory results (Zhang et al., 2009; Zhou et al., 2019), as the number of data is not large in movie

* Corresponding author.
E-mail address: sangjae@sejong.ac.kr (S. Lee).

https://doi.org/10.1016/j.heliyon.2020.e04260
Received 6 February 2020; Received in revised form 7 April 2020; Accepted 16 June 2020

dataset and the number of eWOM and movie related variables is not small, we focus on other non-deep learning methods such as decision trees and k-NN besides neural networks. We employ decision trees and k-NN as these are well applied methods in box office prediction (Kim et al., 2015; Liu et al., 2016; Zhou et al., 2019), and based on the line of researches, this study compares decision trees using ensemble methods (random forests, bagging and boosting) with the k-nearest-neighbors using ensemble methods (bagging and boosting) and linear regression using ensemble methods (bagging and boosting) for predicting movie box office revenue. Although research on an ensemble have been applied to many studies, i.e. online user reviews, customer reviews, product reviews and sales in e-commerce, social networking, multi-industry bankruptcy prediction etc., there is no any research with ensemble prediction combining individual models or averaging these models into one platform for the better prediction of movie box office revenue. This study intends to fill this gap. This study intends to investigate using complex methods of ensemble methods to collect data, by combining and comparing with decision trees using ensemble methods such as random forests, k-NN, linear regression to verify performance of prediction using two well-known techniques: 1) bagging known as bootstrap aggregation which generates multiple random data samples and sampling these data with replacement from the original data; 2) boosting which improves area in the data where model makes errors.

Second, as previous studies regarding the application effectiveness of ensemble methods are lacking, this intends to fill this gap by comparing the prediction performance between ensemble methods and non-ensemble methods within each algorithm. In order to prove the effectiveness of the ensemble, we compares decision trees using ensemble methods (random forests, bagging, boosting), k-NN using ensemble methods (bagging and boosting) and linear regression using ensemble methods (bagging and boosting) with decision trees, k-NN and linear regression using non-ensemble methods.

Third, previous studies have mainly investigated US movies or movies from other countries, and this study fill this gap by utilizing Korean movie data, which has not been examined sufficiently in the movie literature. This study adopts Korean movie data for the movie prediction using ensemble methods.

2. Theoretical background

2.1. Box office revenue prediction

Box office revenue prediction is important to movie producers and directors as their real source of income. In order to predict accurately revenue, it is of a vital importance how to utilize a large amount of data generated from movies produced every year in different countries. There are numerous studies using eWOM, reviews, blog, posts based on social media contents in the past to predict movies box office revenue. Social media based prediction is mostly used for prediction. For example, Asur and Huberman (2010) used social media contents to predict movie box office revenue, and prove that sentiments obtained from the Twitter can be further used to enhance the predicting power. Micro blog and online social networks based prediction can make use of customer reviews and sentiment analysis etc (Liu.et al., 2016). eWOM increase consumer awareness about the movie (Qin, 2001) plays a vital role for the prediction of movie and currently data are mostly collected with user reviews, online user ratings, surveys and interviews (Kim et al., 2015). Online reviews, video sites, blogs, social networks are the forms of eWOM and affect the box office performance (Kim et al., 2013). In the same way, Chintagunta et al. (2010) found that the valence of eWOM (i.e. user rating) has an important role and positive effect on box office prediction.

Nonlinear models are used for better and accurate prediction like Bayesian belief network (BBN) known as the causal belief network is designed to examine the causal relationship between various movie attributes and sensitivity analysis is used to perform prediction of box office success (Lee and Chang, 2009). Back propagation neural network is robust performance and eradicates the limitations of the regression method, more reliable and achieve satisfactory results and useful to solve the problem (Zhang et al., 2009). For instance, a convolutional neural network (CNN) is suggested to obtain features from movie posters, and assess the performance of the proposed multimodal deep neural network, with other prediction methods (Zhou et al., 2019). Given the number of data is not such sufficient and the large number of variables, we focus on other non-deep learning methods such as decision trees and k-NN besides neural networks. Decision support systems (DSS) used to support complex decision-making and problem solving tasks using business Intelligence systems (Delen et al., 2007). As a decision support method based on business intelligence, decision trees and k-NN have been well utilized in box office prediction (Kim et al., 2015; Liu et al., 2016; Zhou et al., 2019) Based on these widely used methods, we proposed ensemble methods in order to investigate the effectiveness of ensemble methods.

2.2. Ensemble prediction

Ensemble methods combines several models to produce better predictive performance rather utilizing a single model (Hansen and Salamon, 1990). Ensemble methods combine multiple supervised models into a “supermodel”. The purpose of using ensemble is to improve the weak power of individual models for the best performance of the combined model to achieve improved predictive accuracy). It consists of a number of learners known as base learners, which can be decision trees, neural networks or other kinds of algorithms. Bagging (Breiman, 1996) and boosting (Freund and Schapire, 1996) are two relatively new but popular methods for producing ensembles with the same algorithm (Bauer and Kohavi, 1999; Drucker and Cortes, 1996; Freund and Schapire, 1996; Quinlan, 1996). Ensemble methods have been used in various applications, bioinformatics problems (gene expression, regulatory elements of DNA and protein sequences) (Yang et al., 2010), bankruptcy prediction (Lee and Choi, 2013), time series prediction (McNames et al., 1999), wind and solar power forecasting (Ren et al., 2015), climate forecasting (Wendy, 2010). Ensemble methods combine several methods in order to lower the prediction error and to achieve higher prediction performance results.

3. Methods

Previous studies on box office revenue prediction which employ eWOM as explanatory variables are limited in that they utilize the forecasting of individual methods and are lacking in combining ensemble methods for prediction of movie box office revenue. Thus, in order to fill this void, our study suggests decision trees using ensemble methods for the prediction of movie box office revenue, and compare k-NN and linear regression using ensemble methods.

This study adopts these three BI methods because this study intends to use both non-statistical and statistical BI methods to examine how they are different in prediction performance, and the methods of combining classifications which provide an operational and fast way to obtain better solutions to optimize predictive measure (Shmueli et al., 2016). These methods are used to examine the difference of prediction performance for box office revenue.

This study constructs three different methods: 1) decision trees using ensemble methods (random forest, bagging, boosting); 2) k-NN using ensemble methods (bagging, boosting); 3) linear regression. For decision trees, k-nearest-neighbors methods, and regression analysis using ensemble methods, 50 different learner models are combined to suggest the prediction of box office revenue. This option controls the number of "weak" classification models that will be created. The ensemble methods will stop when the number or classification models created reaches the value set for this option.

Decision trees method is one of the most popular techniques of a simple structure where terminal nodes follow decision outcomes and

non-terminal nodes show tests on one or more attributes. This study does not set the limit on tree growth or conditions in pruning trees. For decision trees using ensemble methods, however, 139 minimum records are set for each terminal node. Random forests are one of the most powerful, fully automated, ensemble methods that leverages the power of many decision trees, judicious randomization, and ensemble learning to produce astonishingly accurate predictive models (Brieman, 2001). They are collections of decision trees that together produce predictions and deep insights into the structure of data (Guo et al., 2015).

Bagging, short for "bootstrap aggregating" is another form of an ensemble methods based on averaging across multiple random data samples which generates multiple random samples (by sampling with replacement from the original data) and the method is termed as "bootstrap sampling" and runs an algorithm on each sample and producing scores. Shmueli et al. (2016) suggests that bagging improves the performance stability of a model and helps avoid over fitting by separately modeling different data samples and then combining the result, so especially useful for algorithms such as regression trees and neural networks.

Boosting is a slightly different method to create ensembles by directly improving areas in the data where model makes errors forcing the model to pay more attention to the records (Zhou, 2012). The steps in boosting include fitting a model to the data, drawing a sample from the data so that misclassified observations (or observations with large prediction errors) have higher probabilities of selection, fitting the model to the new sample and repetition of drawing a sample from the data, fitting the model to the new sample multiple times.

k-NN is a machine-learning algorithm, which is simple and easy to use which stores all available cases and classifies new cases depend on similarity measure. The algorithm stores the training sample and in new data object it finds the k data objects in the training sample that are closest to it, and only tuning parameter of k-NN is k (Khalid et al., 2013). It is more generic and robust on real world data. Another characteristic of k-NN is it is lazy, which need not to do any generalization from training sample. It has no training phase but more computation in testing phase. This study uses best-k method as the baseline model in k-NN which shows the prediction with the lowest error in validation sample among different k. k-NN using bagging and boosting is compared with k-NN using best k method. For k-NN prediction, the maximum number of nearest neighbors (k) is set to be 20 and 'best-k' is shown within 20 nearest neighbors according to error in validation sample.

The assessment of models performance is proceeded by data partition and providing prediction errors in validation sample after the whole data sample is divided into training and validation sample. Training sample is the largest partition used to learn parameters or weights in the suggested multiple models. Validation sample is used to determine the performance of each learned model from training sample to select the best model after comparing multiple models. Our dataset consists of 13 variables, which are selected as the input variables, and revenue at week 1, 2, 3 as output variables. The prediction errors in validation sample are provided in terms of root mean square error (RMSE). The tool used for this experiment is the XLMiner 2016 version for the predictive analysis with supervised learning algorithms to experiment using different ensemble methods. 40 movies are partitioned as validation sample while the remaining 1399 movies as training sample. This study begins with the decision trees using ensemble methods using random forests, bagging, boosting in a sequence to predict the revenue at the week 1, week 2 and week 3 after release, respectively. We uses 36 fold cross-validation partition samples where in the beginning, from record #1 to record #40 are used as validation sample and the remaining data are used as training sample. Ensemble methods provide a prediction on the target variables, and these predicted values are compared with the actual value in the validation sample. The validation samples are replaced with the next in 36 folds and the process is repeated for 36 times to produce 36 errors from 36 non-overlapping validation samples The aggregated forecasting performance for the revenue at week 1, 2, 3 are presented.

Figure 1 shows the procedure for the comparison of ensemble methods for movie box office revenue prediction.

4. Data collection

In this study, we collected eWOM data from Naver movies site (http://movie.naver.com) targeting the movies released between January 2014 and May 2016. The web crawling method is used to collect data from NAVER site, which is a number one search engine in Korea. This study is using 1439 movies for the prediction of box office revenues, which is composed of 1399 movies as a training sample and 40 movies as a validation data. Data are collected in 2016 after the release of the movies. For the chosen motion pictures, SNS related movie specific and ticket sale of box office data are collected between one week ahead of release and three weeks after release. The study chooses eWOM and movie related variables as independent variables which previous literature suggested as important. This study suggests three output variables: the weekly box office revenue for the first, second and third week after the release of the movie. Models such as decision trees using ensemble methods (random forests, bagging, boosting), k-NN using ensemble methods (bagging, boosting), and linear regression using ensemble methods (bagging, boosting) have been built to predict the movie box office revenue. To predict box office revenue as a dependent variable, five eWOM variables and eight movie related variables are used as independent variables. The variables used in this study are summarized in Table 1. The descriptive statistics of samples are presented in Table 2 describing age, genre, award, nation, sequel, and timing release. The movies which are for teenagers, drama genre, not awarded, released in South Korea, having no sequel, released in holidays starting are having the greater proportion than the others in the sample. The movies which are for teenagers, family movie genre, awarded, Chinese movie, sequel movie, released not in holidays starting are having greater revenue than the others.

5. Results and discussions

This research applies the ensemble methods based on non-statistical and statistical models for the prediction of box office revenue. Decision trees and k-NN are used as non-statistical methods. Linear regression is used as a statistical method on a selected movie dataset that fits a linear model of the form. This study uses ensemble methods (random forests, bagging, boosting) applied to decision trees to improve the forecasting efficiency. The decision trees method using ensemble methods is compared with k-NN and linear regression using ensemble methods. Individual forecasts using three decision trees methods are combined and averaged, and compared with k-NN ensemble methods and linear regression ensemble methods. The predicted values are combined to achieve a single prediction of the combination model. This study consists of several steps, which includes data collection, variable selection of model, compute average of decision trees, k-NN, linear regression using ensemble methods to compare prediction accuracy of each methods.

To predict the box office revenue for a week t, we use eWOM variables at week t-1. Each model uses five eWOM variables (including average review rating, average number of reviews, average number of emotional reviews, helpfulness, total helpful votes review) and eight movie related variables as independent variables. Also, award, film rating, sequel, timing of release, genre, and nationality (variables representing South Korea and USA) are the control variables. 1439 movies are included for predicting box office revenue at week 1, 2 and 3 with ensemble methods analysis.

For each experiment in 36 fold cross-validation, the prediction value is compared with actual value for each movie in validation sample and RMSE is provided. The data for RMSE are provided for dependent variables, i.e., revenue at week 1, 2, 3, and compared among different ensemble methods. Furthermore, we combined all three decision trees using ensemble methods (random forests, bagging and boosting) and

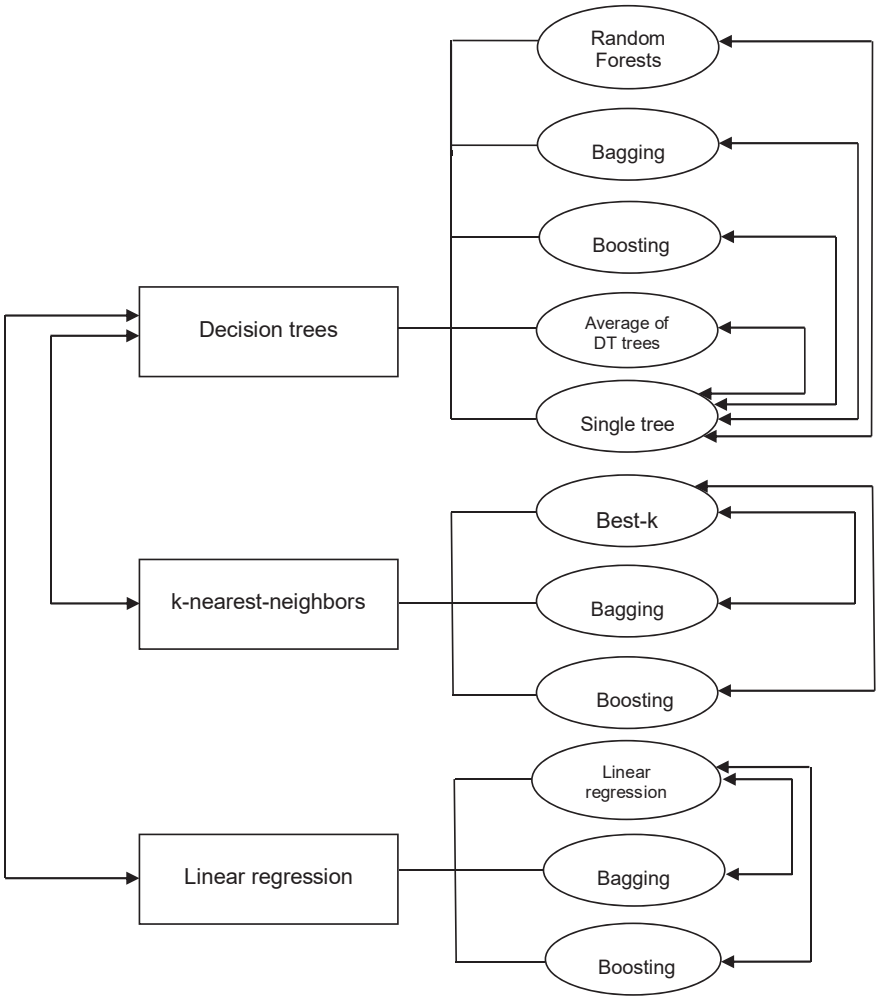

Figure 1. Comparison of ensemble methods for movie box office revenue prediction.

compare this average with k-NN and linear regression. Further, the results of t-test to compare average errors between a pair of methods. Tables 3, 4, and 5 shows the comparison results between decision trees (random forests, bagging, boosting) and k-NN, while Tables 6, 7, and 8 represents the comparison results between decision trees and linear regression.

The results in Tables 3, 4, and 5 show that decision trees using ensemble methods (random forests, bagging, boosting) outperform k-NN using best-k and ensemble methods (bagging, boosting), as the differences of RMSE are highly significant. Similarly, decision trees using ensemble methods (random forests, bagging, boosting) results have lower prediction performance when comparing with k-NN using ensemble methods (bagging, boosting) (Table 4 & Table 5). Similarly, each decision trees using ensemble methods (random forests, bagging, boosting) and their average is compared with linear regression. Decision trees performs better than linear regression in predicting the revenue at

Table 1. Summary of variables.

Category	Variables	Description	Number of possible values
eWOM	Average review rating	Represents average of review rating.	Real values
	Average number of reviews	Represents the average number of user reviews until the movie is released in a new market	Real values
	Average Emotional reviews	Represents the proportion of emotional reviews among total reviews	Real values
	Helpfulness	Proportion of positive answers to total answers to question asking if the review is helpful	Real values
	Total helpfulness votes for reviewer	Represents the total number of helpfulness votes for reviewer	Real values
Movie related variables	Award	Indicates whether a movie got awards winners/nominees (value of 1) or not (value of 0)	2
	Code film rating	Rating of a movies (2 stars, 3stars, 5 stars)	Real values
	Sequel	Indicates whether a movie is a sequel (value of 1) or not (value of 0).	2
	Timing of release	Indicates whether the time of release is high (peak season) and low season	2
	Similar movie revenue	Revenue of competing movies release in first day (d1), first week (wk1), second week (wk2)	Real values
	Genre	Represents the content category the movie belongs to. A movie can be belonged to more than one content category at the same. This study chooses one dummy variable for the drama genre to which our sample belongs to in the greatest proportions.	2
	Nationality (Nation1,Nation2)	Movie released in the respective country (South Korea, USA)	Real values

Table 2. Descriptive statistics of samples.

	Frequency	Percent	Mean (revenue sum) (Won)
a) Age			
No restrictions on age	229	15.9	1.838E+09
Allowed for teenager (12 < age<18)	815	56.5	3.046E+09
Allowed for Adults (age>18)	395	27.4	1.596E+09
Total	1439	100	2.456E+09
b) Genre			
Action	101	7	1.16E+07
Animation	95	6.6	2.09E+09
Comedy	176	12	2.24E+09
Crime	22	1.5	5.38E+08
Documentary	89	7	1.65E+10
Drama	673	46.8	2.02E+10
Family	5	0.3	7.62E+08
Fantasy	4	0.3	6.57E+09
History	2	0.1	3.99E+09
Horror	54	3.8	2.47E+09
Music, Romance	147	10	1.21E+09
Musical	5	0.3	3.03E+08
Science Fiction	8	0.5	5.83E+09
Thriller, Mystery	55	3.8	2.22E+09
War	1	0	2.12E+09
Total	1439	100	9.22E+07
c) Award			
Not awarded	1412	98.1	2.315E+09
Awarded	27	1.9	9.840E+09
Total	1439	100	2.456E+09
d) Nation			
South Korea	571	40	2.738E+09
USA	394	27	2.769E+09
UK	66	4.6	3.504E+09
France	87	6	1.924E+09
Japan	127	8.8	1.138E+09
China	13	1	5.213E+09
Others	181	12.6	1.484E+09
Total	1439	100	2.456E+09
e) Sequel			
No sequel	1415	98.3	2.315E+09
Sequel	24	1.7	9.840E+09
Total	1439	100	2.456E+09
f) Timing release			
Released in holidays starting	907	63.0	1.902E+09
Released in other times	532	37	3.401E+09
Total	1439	100	2.456E+09

week 1 only and when the average of three ensemble methods are combined. This indicates that decision trees using ensemble methods are better in predicting movie revenue than k-NN and are slightly better than linear regression, and especially outperform linear regression in predicting revenue at week 1.

This study further investigates whether decision trees, k-NN, and linear regression provide better prediction performance when they are using ensemble methods, or testing the 'application effectiveness' of ensemble methods. This study compares decision trees using ensemble methods (random forests, bagging, boosting) with single decision tree using non-ensemble methods (Table 9). Similar comparisons are performed for k-NN and linear regression (Tables 10 and 11). Decision trees using ensemble methods (random forest, bagging, and boosting) outperform single decision trees (Table 9). This states that decision trees using ensemble methods is far better tools in predicting box office revenue than single tree using non-ensemble methods, showing much application effectiveness for ensemble methods. k-NN using ensemble methods (bagging, boosting) are also compared with k-NN (best-k) and the result shows that ensemble methods have lower prediction performance than k-NN using the best k method (Table 10). Linear regression using ensemble methods (bagging and boosting) are compared with linear regression alone and ensemble methods show no difference from non-ensemble methods (Table 11). This shows that while decision trees have high application effectiveness for ensemble methods, k-NN and linear regression show low application effectiveness for ensemble methods. This explains that decision trees using ensemble methods outperform k-NN (and partially outperform linear regression).

The results suggests that difference in prediction performance in ensemble methods exist and decision tree using ensemble methods (random forests, bagging, boosting) perform better than k-NN using ensemble methods (bagging, boosting), and linear regression (bagging, boosting) in week 1. In summary, this study suggests that decision trees using ensemble methods (random forests, bagging, boosting) and their average of ensemble methods can potentially lead to smaller prediction errors, and therefore provide a better predictive power. The reason can be explained by that k-NN and linear regression using ensemble methods have lower application effectiveness than decision trees using ensemble methods.

6. Conclusion and implications

The study provide insights to the literature on movie revenue prediction and practitioners in movie industry. First, as the studies have been lacking on ensemble prediction combining individual models for the better prediction of movie box office revenue, our study intends to fill this gap by suggesting ensemble methods in predicting box office revenue provide implications to researchers in movie revenue prediction. While many researches exist on the prediction of movies using social

Table 3. Comparing decision trees using ensemble methods with k-nearest-neighbors (best-k).

Compared Models		Mean	T	Sig. (2-tailed)
Week 1	decision trees (random forest) — k-nearest-neighbors	-4.821E+08	-3.874	0.000
	decision trees (bagging) — k-nearest-neighbors	-4.834E+08	-3.610	0.001
	decision trees (boosting) — k-nearest-neighbors	-5.131E+08	-3.224	0.003
	decision trees (random forest, bagging, boosting) — k-nearest-neighbors	-5.941E+08	-4.332	0.000
Week 2	decision trees (random forest) — k-nearest-neighbors	-5.174E+08	-5.710	0.000
	decision trees (bagging) — k-nearest-neighbors	-4.976E+08	-5.147	0.000
	decision trees (boosting) — k-nearest-neighbors	-3.563E+08	-2.643	0.012
	decision trees (random forest, bagging, boosting) — k-nearest-neighbors	-5.085E+08	-5.004	0.000
Week 3	decision trees (random forest) — k-nearest-neighbors	-2.841E+08	-4.993	0.000
	decision trees (bagging) — k-nearest-neighbors	-2.733E+08	-3.462	0.001
	decision trees (boosting) — k-nearest-neighbors	-1.798E+08	-1.938	0.061
	decision trees (random forest, bagging, boosting) — k-nearest-neighbors	-2.788E+08	-3.859	0.000

Table 4. Comparing decision trees using ensemble methods with k-nearest-neighbors (bagging).

Compared Models		Mean	t	Sig. (2-tailed)
Week 1	decision trees (random forest) — k-nearest-neighbors (bagging)	-8.505E+08	-7.561	0.000
	decision trees (bagging) — k-nearest-neighbors (bagging)	-8.519E+08	-7.185	0.000
	decision trees (boosting) — k-nearest-neighbors (bagging)	-8.816E+08	-6.289	0.000
	decision trees (random forest, bagging, boosting) — k-nearest-neighbors (bagging)	-9.626E+08	-8.068	0.000
Week 2	decision trees (random forest) — k-nearest-neighbors (bagging)	-8.618E+08	-7.869	0.000
	decision trees (bagging) — k-nearest-neighbors (bagging)	-8.420E+08	-7.301	0.000
	decision trees (boosting) — k-nearest-neighbors (bagging)	-7.007E+08	-4.607	0.000
	decision trees (random forest, bagging, boosting) — k-nearest-neighbors (bagging)	-8.529E+08	-7.051	0.000
Week 3	decision trees (random forest) — k-nearest-neighbors (bagging)	-5.638E+08	-7.751	0.000
	decision trees (bagging) — k-nearest-neighbors (bagging)	-5.530E+08	-6.229	0.000
	decision trees (boosting) — k-nearest-neighbors (bagging)	-4.596E+08	-4.649	0.000
	decision trees (random forest, bagging, boosting) — k-nearest-neighbors (bagging)	-5.585E+08	-6.736	0.000

Table 5. Comparing decision trees using ensemble methods with k-nearest-neighbors (boosting).

Compared Models		Mean	t	Sig. (2-tailed)
Week 1	decision trees (random forest) — k-nearest-neighbors (boosting)	-1.063E+09	-7.698	0.000
	decision trees (bagging) — k-nearest-neighbors (boosting)	-1.064E+09	-7.456	0.000
	decision trees (boosting) — k-nearest-neighbors (boosting)	-1.094E+09	-7.101	0.000
	decision trees (random forest, bagging, boosting) — k-nearest-neighbors (boosting)	-1.175E+09	-8.496	0.000
Week 2	decision trees (random forest) — k-nearest-neighbors (boosting)	-9.864E+08	-6.373	0.000
	decision trees (bagging) — k-nearest-neighbors (boosting)	-9.667E+08	-6.165	0.000
	decision trees (boosting) — k-nearest-neighbors (boosting)	-8.254E+08	-4.504	0.000
	decision trees (random forest, bagging, boosting) — k-nearest-neighbors (boosting)	-9.776E+08	-6.060	0.000
Week 3	decision trees (random forest) — k-nearest-neighbors (boosting)	-6.168E+08	-6.211	0.000
	decision trees (bagging) — k-nearest-neighbors (boosting)	-6.060E+08	-5.190	0.000
	decision trees (boosting) — k-nearest-neighbors (boosting)	-5.126E+08	-4.073	0.000
	decision trees (random forest, bagging, boosting) — k-nearest-neighbors (boosting)	-6.115E+08	-5.512	0.000

Table 6. Comparing decision trees using ensemble methods with linear regression.

Compared Models		Mean	t	Sig. (2-tailed)
Week 1	decision trees (random forest) — linear regression	-1.842E+08	-1.367	0.180
	decision trees (bagging) — linear regression	-1.856E+08	-1.328	0.193
	decision trees (boosting) — linear regression	-2.153E+08	-1.606	0.117
	decision trees (random forest, bagging, boosting) — linear regression	-2.963E+08	-2.303	0.027
Week 2	decision trees (random forest) — linear regression	-6.204E+07	-.850	0.401
	decision trees (bagging) — linear regression	-4.228E+07	-.555	0.582
	decision trees (boosting) — linear regression	9.901E+07	1.315	0.197
	decision trees (random forest, bagging, boosting) — linear regression	-5.318E+07	-.811	0.423
Week 3	decision trees (random forest) — linear regression	3.689E+06	.076	0.940
	decision trees (bagging) — linear regression	1.451E+07	.392	0.697
	decision trees (boosting) — linear regression	1.080E+08	2.100	0.043
	decision trees (random forest, bagging, boosting) — linear regression	9.034E+06	.250	0.804

networking, blogs, and linear regression methods, there is almost no prior research of decision trees using ensemble methods in box office revenue prediction. This study compares decision trees using ensemble methods (random forests, bagging and boosting) with the k-NN using ensemble methods (bagging and boosting) and linear regression using ensemble methods (bagging and boosting) for predicting movie box office revenue. Decision trees using ensemble methods have greater prediction performance for the first week, second week and third week respectively in the box office revenue than k-NN and linear regression after analyzing 1439 movies.

Second, as previous studies have not investigated the application effectiveness of ensemble methods of business intelligence methods, this study purports to fill this gap by comparing the prediction performance between ensemble methods and non-ensemble methods within each algorithm. For decision tree methods, unlike the other methods, the prediction performance of ensemble methods is greater than that of non-ensemble methods. This shows that decision trees using ensemble methods provide better application effectiveness of ensemble methods than k-NN and linear regression analysis.

Third, this study analyzes Korean movie data, which have been rarely investigated sufficiently in the movie literature. Previous studies have mainly centered on using US movies or movies from other countries. This study show the prediction results using Korean movie data for the movie prediction, providing the results using international movie markets, which may be of interests to practitioners working in global movie industry.

Table 7. Comparing decision trees using ensemble methods with linear regression (bagging).

Compared Models		Mean	t	Sig. (2-tailed)
Week 1	decision trees (random forest) — linear regression (bagging)	-1.802E+08	-1.257	0.217
	decision trees (bagging) — linear regression (bagging)	-1.816E+08	-1.230	0.227
	decision trees (boosting) — linear regression (bagging)	-2.112E+08	-1.489	0.145
	decision trees (random forest, bagging, boosting) — linear regression (bagging)	-2.923E+08	-2.131	0.040
Week 2	decision trees (random forest) — linear regression (bagging)	-7.242E+07	-1.038	0.306
	decision trees (bagging) — linear regression (bagging)	-5.267E+07	-.757	0.454
	decision trees (boosting) — linear regression (bagging)	8.863E+07	1.223	0.229
	decision trees (random forest, bagging, boosting) — linear regression (bagging)	-6.357E+07	-1.046	0.303
Week 3	decision trees (random forest) — linear regression (bagging)	-7.922E+06	-.148	0.883
	decision trees (bagging) — linear regression (bagging)	2.899E+06	.082	0.935
	decision trees (boosting) — linear regression (bagging)	9.634E+07	1.899	0.066
	decision trees (random forest, bagging, boosting) — linear regression (bagging)	-2.577E+06	-.068	0.946

Table 8. Comparing decision trees using ensemble methods with linear regression (boosting).

Compared Models		Mean	t	Sig. (2-tailed)
Week 1	decision trees (random forest) — linear regression (boosting)	-1.806E+08	-1.434	0.160
	decision trees (bagging) — linear regression (boosting)	-1.819E+08	-1.391	0.173
	decision trees (boosting) — linear regression (boosting)	-2.116E+08	-1.663	0.105
	decision trees (random forest, bagging, boosting) — linear regression (boosting)	-2.926E+08	-2.435	0.020
Week 2	decision trees (random forest) — linear regression (boosting)	-5.891E+07	-.932	0.358
	decision trees (bagging) — linear regression (boosting)	-3.916E+07	-.599	0.553
	decision trees (boosting) — linear regression (boosting)	1.021E+08	1.415	0.166
	decision trees (random forest, bagging, boosting) — linear regression (boosting)	-5.006E+07	-.891	0.379
Week 3	decision trees (random forest) — linear regression (boosting)	1.785E+05	.004	0.997
	decision trees (bagging) — linear regression (boosting)	1.100E+07	.325	0.747
	decision trees (boosting) — linear regression (boosting)	1.044E+08	2.034	0.050
	decision trees (random forest, bagging, boosting) — linear regression (boosting)	5.523E+06	.162	0.872

Table 9. Comparing decision trees using ensemble methods with single decision trees.

Compared Models		Mean	t	Sig. (2-tailed)
Week 1	decision trees (random forest) — single decision trees	-1.110E+08	-2.547	0.015
	decision trees (bagging) — single decision trees	-1.123E+08	-2.906	0.006
	decision trees (boosting) — single decision trees	-1.420E+08	-1.607	0.117
	decision trees (random forest, bagging, boosting) — single decision trees	-2.230E+08	-4.335	0.000
Week 2	decision trees (random forest) — single decision trees	-1.497E+08	-3.077	0.004
	decision trees (bagging) — single decision trees	-1.299E+08	-2.527	0.016
	decision trees (boosting) — single decision trees	1.139E+07	.112	0.911
	decision trees (random forest, bagging, boosting) — single decision trees	-1.408E+08	-2.421	0.021
Week 3	decision trees (random forest) — single decision trees	-1.575E+07	-.560	0.579
	decision trees (bagging) — single decision trees	-4.930E+06	-.177	0.861
	decision trees (boosting) — single decision trees	8.852E+07	1.712	0.096
	decision trees (random forest, bagging, boosting) — single decision trees	-1.041E+07	-.389	0.699

Table 10. Comparing k-nearest-neighbors using ensemble methods with k-nearest-neighbors (best-k).

Compared Models		Mean	T	Sig. (2-tailed)
Week 1	k-nearest-neighbors (bagging) — k-nearest-neighbors (best-k)	5.702E+08	4.214	0.000
	k-nearest-neighbors (boosting) — k-nearest-neighbors (best-k)	7.823E+08	4.897	0.000
	k-nearest-neighbors (bagging, boosting) — k-nearest-neighbors (best-k)	6.421E+08	4.454	0.000
Week 2	k-nearest-neighbors (bagging) — k-nearest-neighbors (best-k)	1.012E+09	7.719	0.000
	k-nearest-neighbors (boosting) — k-nearest-neighbors (best-k)	1.136E+09	6.401	0.000
	k-nearest-neighbors (bagging, boosting) — k-nearest-neighbors (best-k)	1.032E+09	6.739	0.000
Week 3	k-nearest-neighbors (bagging) — k-nearest-neighbors (best-k)	-2.915E+08	-1.862	0.071
	k-nearest-neighbors (boosting) — k-nearest-neighbors (best-k)	-2.385E+08	-1.367	0.180
	k-nearest-neighbors (bagging, boosting) —k-nearest-neighbors (best-k)	-3.066E+08	-1.874	0.069

Table 11. Comparing linear regression using ensemble methods with linear regression.

Compared Models		Mean	t	Sig. (2-tailed)
Week 1	linear regression (bagging) — linear regression	-3.672E+06	-.219	0.828
	linear regression (boosting) — linear regression	-4.016E+06	-.187	0.852
	linear regression (bagging, boosting) — linear regression	-3.844E+06	-.237	0.814
Week 2	linear regression (bagging) — linear regression	-3.122E+06	-.114	0.910
	linear regression (boosting) — linear regression	1.039E+07	.285	0.777
	linear regression (bagging, boosting) — linear regression	3.632E+06	.120	0.905
Week 3	linear regression (bagging) — linear regression	3.510E+06	.383	0.704
	linear regression (boosting) — linear regression	1.161E+07	.688	0.496
	linear regression (bagging, boosting) — linear regression	7.561E+06	.686	0.497

There are limitations and future research issues. The small number of data may explain the reason that k-NN (best-k) has lower prediction errors than k-NN using bagging and boosting. We admit the lack of generalizability of our results. Future research could further test k-NN (bagging, boosting) to provide the robustness of prediction results using a large group of movies data. In addition, the results of our research are limited in that they are based on the Korean movie data from the Naver portal site only. The movie sample from more American, or other countries should be considered in the future research. Finally, the methods presented in this study can be tested with other algorithms using ensemble methods such as neural network which can perform better using a larger number of movie data. Further, the future study can reveal whether ensemble methods combining several algorithms to predict box office result in greater prediction performance. Further ensemble methods of this study which is a little general approach in previous studies can be further developed to reflect more unique hybrid approach in terms of variable selection or learning methods with other techniques in order to produce a more differentiated approach than previous ensemble methods.

Declarations

Author contribution statement

S. Lee and B. KC: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

J.Y. Choeh: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Asur, S., Huberman, B.A., 2010. Predicting the future with social media [C]//Web intelligence and intelligent agent technology (WI-IAT). IEEE/WIC/ACM international conference on IEEE 1, 492–499.

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

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. 36, 105–139.

Breiman, L., 1996. Stacked regressions. Mach. Learn. 24 (1), 49–64.

Brieman, L., 2001. In: Machine Learning, 45. Kluwer Academic Publishers, pp. 5–32, 1.

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. Market. Sci. 29 (5), 944–957.

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

Drucker, H., Cortes, C., 1996. In: Boosting Decision Trees, Advances in Neural Information Processing Systems, 8. MIT Press, Cambridge, MA, pp. 479–485.

Freund, Y., Schapire, R., 1996. Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 148–156.

Guille, A., Hacid, H., 2012. A predictive model for the temporal dynamics of information diffusion in online social networks. In: WWW '12 Companion Proceedings of the 21st International Conference Companion on World Wide Web. ACM, NY, pp. 1145–1152.

Guo, Z., Zhang, X., Hou, Y., 2015. Predicting box office receipts of movies with pruned random forest. International Conference on Neural Information Processing 9489, 55–62.

Hansen, L.K., Salamon, P., 1990. Neural networks ensembles. IEEE Trans. Pattern Anal. Mach. Intell. 12 (10), 993–1001.

Khalid, A., Najadat, H., Hmeidi, I., Shatnawi, M., 2013. Stock price prediction using k-nearest-neighbors (k-NN) algorithm. Int. J. Bus. Humanit. Technol. 3 (3).

Kim, S., Park, N., Park, S., 2013. Exploring the effects of online word of mouth and expert reviews on theatrical movies' box office success. J. Media Econ. 114 (2), 26–98.

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

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

Lee, S., Choi, W.S., 2013. A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. Expert Syst. Appl.: Int. J. 40 (8), 2941–2946.

Leenders, M.A.A.M., Eliashberg, J., 2011. The antecedents and consequences of restrictive age-based ratings in the global motion picture industry". Int. J. Res. Market. 28, 367–377.

Liu, T., Ding, X., Chen, Y., Chen, H., Guo, M., 2016. Predicting movie box office revenues by exploiting large-scale social media content. J. Multimed. Tool. Appl. 75 (3), 1509–1528.

McNames, J., Suykens, J., Vandewalle, J., 1999. Winning entry of the K.U. Leuven time series prediction competition. Int. J. Bifurcat. Chaos 9 (8), 1485–1500.

Oh, C., Roumani, Y., Nwankpa, J.K., Hu, H.-F., 2017. Beyond likes and tweets: consumer engagement behavior and movie box office in social media. Inf. Manag. 4 (Issue 1), 25–37.

Qin, L., 2001. Word of Blog for Movies: a predictor and an outcome of box office revenue? J. Electron. Commer. Res. 12 (3).

Quinlan, J.R., 1996. bagging, boosting, and c4.5. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence, pp. 725–730.

Ren, Y., Suganthan, P.N., Srikanth, N., 2015. Ensemble methods for wind and solar power forecasting – a state-of-the-art review. Renew. Sustain. Energy Rev. 50, 82–91.

Ru, Y., Li, B., Liu, J., Chai, J., 2018. An effective daily box office prediction model based on deep neural networks. Cognit. Syst. Res. 52, 182–191.

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

Shmueli, G., Bruce, P., Patel, N., 2016. Data Mining for Business Analytics: Concepts, Techniques, and Applications with XL Miner, Third Edition. John Wiley & Sons, Inc., pp. 292–297

Wang, Z., Zhang, J., Ji, S., Meng, C., Li, T., Zheng, Y., 2020. Predicting and ranking box office revenue of movies based on big data. Inf. Fusion 60, 25–40.

Wendy, S.P., 2010. Predicting weather and climate: uncertainty, ensembles and probability. Studies in History and philosophy of Modern Physics 41, 263–272.

Yang, P., Ho, J.W.K., Zomaya, A.Y., Zhou, B.B., 2010. A genetic ensemble approach for gene-gene interaction identification. BMC Bioinf. 524 (11), 1471–2105.

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

Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC Press, p. 15.

Zhou, Y., Zhang, L., Yi, Z., 2019. Predicting movie box-office revenues using deep neural networks. Neural Comput. Appl. 31, 1855–1865.