ICIC Express Letters
Part B: Applications
Volume **13**, Number **11**, November **2022**
 ISSN 2185-2766
pp. **1207–1214**

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR PREDICTION OF MOVIE'S BOX OFFICE SUCCESS

Jin Young Choi and Gun Ho Lee

Department of Industrial and Information Systems Engineering
Soongsil University
369 Sangdo-ro, Dongjak-gu, Seoul 06978, Korea
hell0u0@soongsil.ac.kr; ghlee@ssu.ac.kr

Received March 2022; accepted June 2022

Abstract. *The film industry is a high-risk, high-return business from an economic point of view. It is necessary to lower the uncertainty of the film industry by predicting the success or failure of the film business before the film's release. This study builds the various models to predict the success of films before the release in order to give helpful information to the film industry in the production stage. After selecting variables, such as director, actor, genre, runtime, grade, language, cost, original story, release month, promotion, and marketing before the movie is released, we visualize and analyze the variables. We implement six predictive models using various algorithms to transform the data to fit the predictive model by performing a pre-data processing process. We compare and evaluate the performances of the predictive models on a variety of criteria, such as predictive accuracy, precision, recall, f1-score, and learning time. We visualize the performance of the models. The deep neural network model shows the highest prediction accuracy, and the random forest model also has relatively high accuracy.*
Keywords: Comparative analysis, Machine learning models, Prediction of movie's box office success, Predictive models, Film business

1. **Introduction.** The film industry is a high-risk, high-yield industry compared to others. The film's success is generally determined around one or two weeks after its release, depending on the good ratings on social media such as Facebook and blogs and the number of releases in theaters. Even after investing a considerable amount of cost from the production stage to the pre-release marketing stage, it often fails to surpass a break-even point. The Break-Even Point (BEP) represents the sales amount in revenue (sales) required to cover total costs, consisting of both fixed and variable costs to the company. In general, the success of a film business is determined by whether it is possible to exceed the break-even point of the return on investment. It is reported that only 27.1% of the films released succeeded in 2015 based on movies released in the Korean film industry [1]. In the film industry, the business officials need to have confidence in the film's success and the audience's interest during the filmmaking and marketing phases. It is necessary to predict the success of a film to operate the film business efficiently. There have been many studies that predict success after a movie is released, but few studies predict success before its release. After the release, movie investors already decided to invest a lot of capital and invested, so the prediction of failure and success could not be of much help to investors. A predictive model can be used as a tool for film investors to decide whether or not to invest in a film. To reduce uncertainty by predicting the success of a film before its release and aid in decision-making, we build a model that predicts the film's success using variables that exist only before its release. In this study, the success or failure of a movie is decided by whether it has crossed the break-even point. We build six predictive models using machine learning techniques to predict whether a movie will be successful or

DOI: 10.24507/icicelb.13.11.1207

not. According to the degree of performance, we compare and analyze six models. After comparing the six models, the model trained by the Deep Neural Network (DNN) had the highest prediction accuracy of 85%.

The remainder of this paper is organized as follows. Section 2 reviews existing studies that used machine learning techniques to make predictions related to movie box office success. In Section 3, the data and dependent and independent variables are introduced. Section 4 describes the machine learning techniques to build predictive models. We compare and analyze the performance of six models in Section 5 and describe the conclusions and future studies derived from the analysis in Section 6.

2. **Related Study.** Recently, the prediction of movie performance has been actively studied using machine learning techniques. Dissanayake and Vidanagama [2] predicted the movie's success using variables such as the number of reviews, the director's Facebook likes, and the actor's Facebook likes. They built a regression model and a classification model. The R2 score of the random forest regression model was 0.68, the best performance, and the SVM (Support Vector Machine) model showed 100% accuracy. Mahmud et al. [3] built a classification model to classify the film's success. The number of Facebook likes, reviews, and faces on posters of directors or actors were used as variables. They built models of the logical regression, SVM, and multi-layer perceptron. They used one class away accuracy to evaluate the models which means the classifier predicted the class of a movie within one class distance. The one class away accuracy of the multi-layer perceptron model showed 85.31%, the best performance. Zhang et al. [4] presented two improved Bass models for predicting box offices for each week through social media analysis using the Naïve Bayes and logistic regression and SVM algorithms. As a result of comparing the performance of the multiple linear regression model, the neural network model, and the improved Bass model, the R2 value of the Bass model was the best, 0.97. Nihalaani et al. [5] attempted to predict the film's success using each of the review datasets labeled positive, negative, and neutral in the Internet Movie Database (IMDb) and metadata sets, including basic information about the film and Facebook-related data. The best algorithm for predicting movie success is logistic regression, with 90% accuracy in the first data set and 100% accuracy in the second dataset. SVM performed second best, followed by Naïve Bayes. Athira and Lakshmi [6] predicted the movie's success using IMDb reviews, YouTube comments, and Twitter posts. Sentiment analysis was performed on each review and comment data, and based on this, they trained the model using the Random Forest algorithm. The accuracy was 86%.

Most of the existing studies focused on prediction using data that are generated after the film is released, such as screen count, reviews, and ratings, which does not provide an advantage to the decision-making of film project managers in the pre-release planning and production stages before it is released. Therefore, this study compares the various models and recommends the best model, which is helpful in decision-making by allowing movie producers and movie industry workers to predict in advance whether the movie will succeed.

3. **Data.** Except for marketing-related data, we collected a total of 11,006 movie data lists from the popular film page and detailed movie data such as rating, genre, release date, runtime, existence of the original, original title, language, cost, and profit from the TMDB (The Movie Database) website. In this study, we use 'runtime', 'existence of the original', 'cost', 'release year', 'release month', 'release date', 'director's reputation', 'actor's reputation', 'average number of actors' participation', 'online marketing', 'rating', 'genre', and 'language' as independent variables. Runtime is a continuous variable that is converted in minutes.

The existence of the original is a variable in whether the film is based on the novel or not. We delete outliers, i.e., data for a short film, non-commercial film, or a film that costs less than $1,000 to produce. We perform min-max scaling prior to model fitting to measure variables at the same scales to contribute equally to the learned function and prevent bias. We calculate the reputation of directors and actors by averaging the net margin (excluding costs from profits) of the work directed or starred by each character. We quantify the director's reputation and the main actors for each film with the average net margin of the director and the average net margin of the main actors (up to five). In addition, we collect the number of participating films for each character. We use the average number of participating films for the director and the leading actor as another variable for the director and actor's reputation. We obtain and use the number of trailer videos and promotional posts posted on the Google site from 180 days before the movie's release. This study categorizes movies as allowed to up to 12 years old, up to 15 years old, not allowed to youth, or allowed to all ages. We consider the number of blog posts that were published in the six months prior to the film's release.

According to the genre, the film consists of 23.2% drama, 20.6% comedy, 18.3% action, 7.8% horror, 7.3% adventure, 5.5% crime, 5.0% thriller, 2.6% SF, 2.6% fantasy, 1.9% romance, and 5.2% others. In order to alleviate this unbalance of the categories, we re-categorize into drama, comedy, action, horror/crime/thriller, adventure/science fiction/fantasy, and others in consideration of the characteristics and proportions of each genre.

As for the original language, it consists of 22 national languages. English accounts for 92.6% of the total, 2.0% French, 1.1% Spanish, 0.6% Italian, 0.6% Korean, 0.5% Chinese, 0.4% Japanese, and 2.2% others. We re-categorize the languages into English, French/Spanish/Italian, Korean/Chinese/Japanese, and Others.

If the cost of each film minus revenue is higher than zero, it is defined as class 1 and class 0 if less than zero. Of the total 4440 movie data, there are 3348 successful movies of class 1 and 1092 unpopular movies of class 0, indicating an imbalance in the data. These data imbalances are likely to make biased predictions for Class 1 that account for a relatively high proportion, which makes it to implement reliable prediction models. This study mitigates the data imbalance by using SMOTE (Synthetic Minority Oversampling TEchnique) and also uses 4692 training sets to train the model and 1332 test sets.

4. **Algorithms.** This study builds six models, i.e., random forest classifier, gradient boosting classifier, SVM classifier, eXtreme Gradient Boosting (XGBoost) classifier, a voting-based ensemble model combined with these four models, and a DNN.

The random forest classifier is an ensemble of a decision tree. As the number of trees increases, the model's generalization errors converge to limits. Generalization errors depend on the strength of individual trees and their correlation. It maintains accuracy even when there is a lot of missing data and produces unbiased estimates [7]. Gradient boosting classifier is the ensemble for a random forest, which builds a strong learner by accumulating a weak learner at each iteration stage. At this point, gradient descent is applied for the loss function at each stage. It is reported that the gradient boosting provides excellent results in terms of accuracy and generalization [7]. SVM is a tool for solving multi-dimensional function estimation problems. After mapping input vectors as high-dimensional feature spaces, the goal is to find a hyperplane that maximizes the margin between classification groups. The advantage is that there are not many kinds of parameters, which are relatively simple to implement and prevent overfitting problems [8]. XGBoost classifier is an improved algorithm based on gradient boosting to construct and operate boosting trees in parallel efficiently. Gradient descent is applied to arbitrary loss functions that measure the model's performance in the training dataset. By using a normalized loss function and

reducing the weights of each tree, we can reduce the impact of a single tree on the final prediction. It also has the advantage of avoiding overfitting [9].

The voting-based classifier is an ensemble technique that combines different machine learning models to predict based on majority predictions, i.e., hard voting or average prediction probabilities, i.e., soft voting. The advantage is that it can compensate for the weaknesses of individual models. This study uses 'soft voting' to assign specific weights to each classifier based on the average predicted probability of the learning models [10]. Neural network is a mathematical model in which neurons are connected to form a network. One neuron calculates the output value by multiplying the weights of several input signals, then summing them and applying them to the activation function. Neural networks are divided into single-layer and multi-layer depending on the number of layers. Single-layer consists of an input layer and an output layer, and a multi-layer has one or more hidden layers between the input and output layers [11]. DNN has been attracting attention as a key model for deep learning for large-scale data due to recent computer performance improvements. DNN can learn a variety of nonlinear relationships with multiple hidden layers [12]. This study constructs a DNN model with three hidden layers for analysis.

5. **Analysis of Models.** This study compares the precision, recall, f1-score, and accuracy values of the six models to evaluate the performance of the models learned. To evaluate the generalization errors of the models and verify the model's excellence, we implement k-fold cross-validation. Evaluating the performance on a training dataset is called in-sample testing. Verifying the predictive performance on a non-training sample dataset is out-of-sample testing or cross-validation. The k-fold cross-validation divides the dataset into k independent subsets. $k-1$ subsets are used as a training set to train the model, while one subset is used as a test set to evaluate generalization errors. Repeat this k times. Each validation set can be reliably used in evaluating generalization errors because it is not used to train the model [13].

Figure 1 compares precisions according to folds $k = 3$, $k = 4$, and $k = 5$ for k-fold cross-validation on each model. Precision is the proportion of what the model classifies as true is actually true, and it can be expressed as precision, also known as Positive Predictive Value (PPV). The six models maintain relatively constant precision values at $k = 3$, 4, and 5. The average precisions of the k-fold cross-validation of DNN, random forest, voting classifier, and XGBoost models are 0.85, 0.84, 0.82, and 0.78, respectively. SVM and gradient boosting models show relatively low average precisions, 0.69 and 0.68.

Figure 2 compares recall according to folds $k = 3$, $k = 4$, and $k = 5$ of k-fold cross-validation on each model. The recall is the ratio of what is true to what the model predicts to be true. The six predictive models show no significant variation in recall

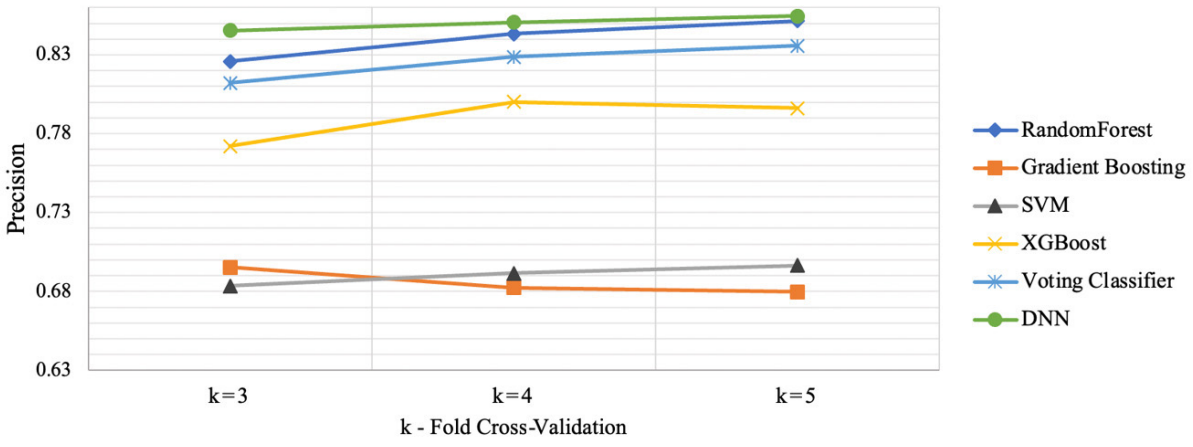

FIGURE 1. Comparison of precision according to fold k

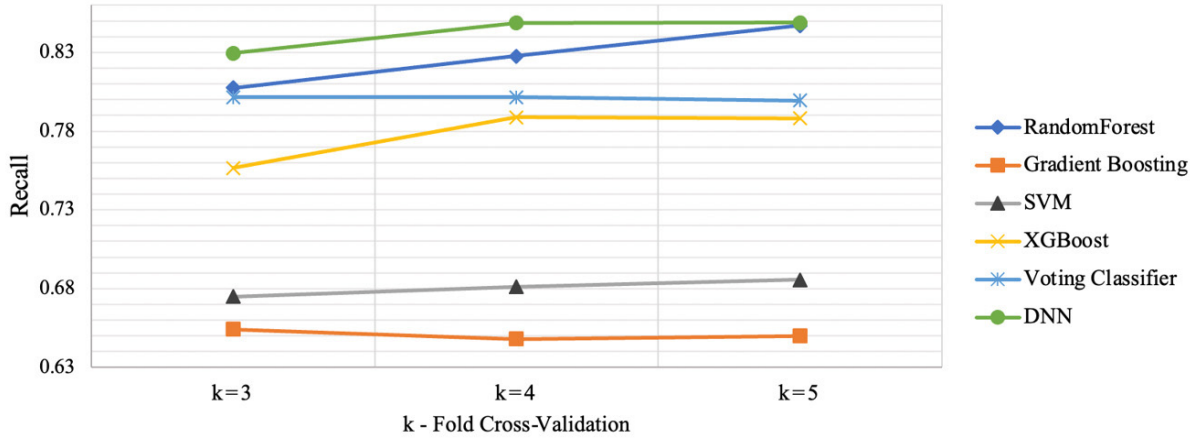

FIGURE 2. Comparison of recall according to fold k

values according to fold k. For folds k = 3, 4, and 5, the highest average recall value is 0.84 for DNN's. The average recalls of the random forest and voting classifier models show relatively high values of 0.82 and 0.80. The average recall values of gradient boosting and SVM models are 0.65 and 0.68.

We evaluate the model's performance through f1-score when the data label is unbalanced. F1-scores of k-fold cross-validation on each model are compared in Figure 3. F1-score is the harmonic mean of precision and recall, which can be expressed as 2(precision × recall)/(precision + recall). The graph shows that the six models do not show a significant variation of f1-score values for the folds. The mean f1-score value of the DNN model is the best, 0.84. The random forest and the voting classifier show relatively high average f1-scores, 0.82 and 0.80. On the other hand, the average f1-score value of the SVM models is 0.67, and the gradient boosting model has the worst f1-score, 0.63.

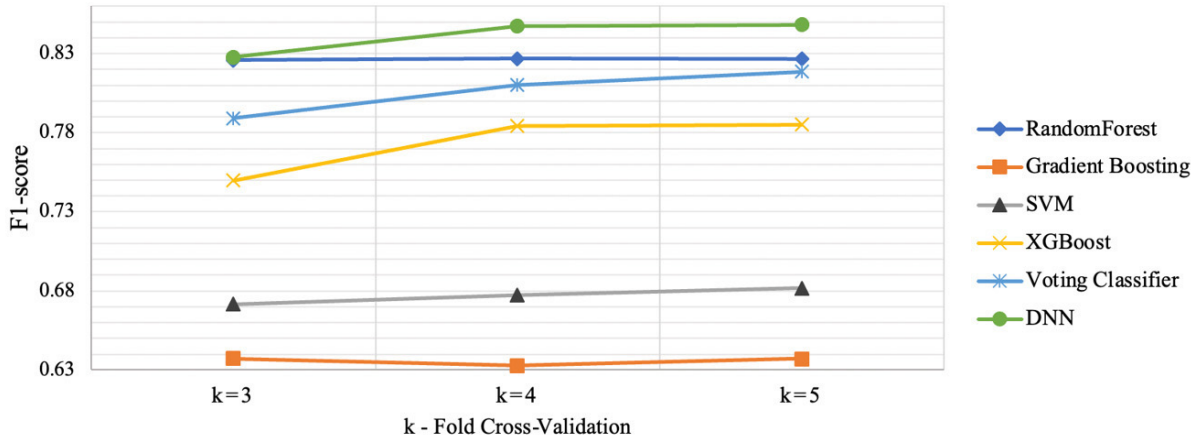

FIGURE 3. Comparison of f1-score according to fold k

Figure 4 compares the average accuracy values after performing five folds of cross-validations on each model. Accuracy is the ratio of accurately predicted cases in the overall case, the evaluation index that can most intuitively represent the model's performance. Accuracy can be expressed as (TP + TN)/(TP + FN + FP + TN). The accuracy of the DNN model shows the highest prediction accuracy of 0.8508. The random forest and voting classifier models have relatively good accuracies of 0.847 and 0.8296. XGBoost, SVM, and gradient boosting show low prediction accuracies of 0.7882, 0.6856, and 0.6501.

Figures 1, 2, and 3 show that the deviation of performances is not high for the various datasets. For four evaluation indicators, precision, recall, f1-score, and accuracy, the DNN

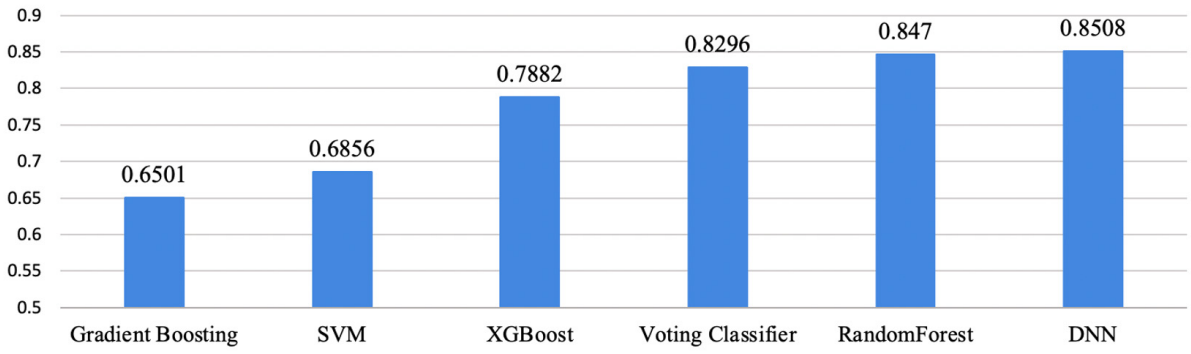

FIGURE 4. Comparison of average accuracy

model is the best, and the gradient boosting model shows the lowest performance. Quader et al. [14], who analyzed historical data of Box Office Mojo and Metacritic to predict the movie's success, used the neural network model and the SVM model as a predictive model. The neural network model showed 84% accuracy based on the pre-release characteristics, and SVM showed 83% accuracy. The neural network model was selected as the best model. In this study, the best model is the DNN with 85% accuracy.

A bar graph in Figure 5 shows the learning time of six models. The XGBoost model completes learning in the shortest time, 0.472 seconds, and random forest, gradient boosting, and SVM model learn models for about a second. The DNN model needs 50.01 seconds to learn, which is 50 seconds longer than the XGBoost model's learning time. The voting classifier model has the second-longest learning, 10.2285 seconds. Evaluated by the training time of the model, XGBoost, random forest, gradient boosting, and SVM model are relatively efficient.

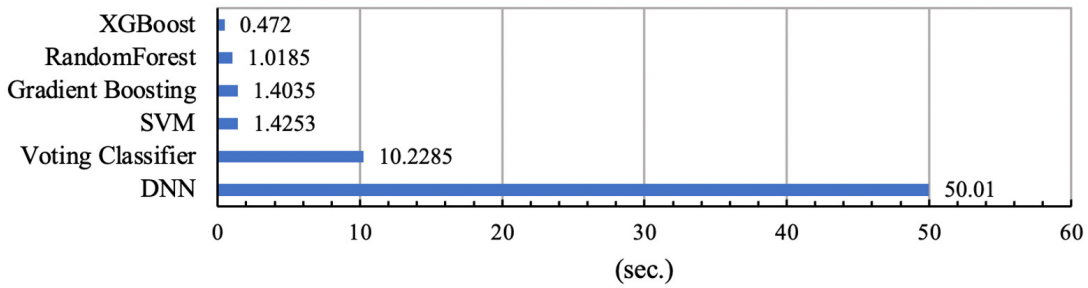

FIGURE 5. Bar chart of learning time (sec.)

The XGBoost model used 18 independent variables for learning. The bar graph in Figure 6 shows the feature importance in the XGBoost. Feature importance represents a degree of relatively how much the variable affects the dependent variable. The larger the value of the feature importance, the greater the influence on the dependent variable. In Figure 6, director_work_cnt and actor_work_cnt are variables that represent the number of works each director or actor previously participated in some movie. The variables of director_margin and act_margin are used for a character's reputation and the average net margin for the previous five works. The sns_marketing is the number of promotional posts posted on Google during the pre-release period, a variable for online marketing. In the XGBoost model, the director's reputation is the essential variable. Four independent variables, including release_month, sns_marketing, cost, and release_year, significantly impact the predictive model. Also, the director_margin is the most critical factor in building predictive models with the random forest and gradient boosting. In other words, the director's reputation has the most significant impact on the dependent variable, i.e., whether it

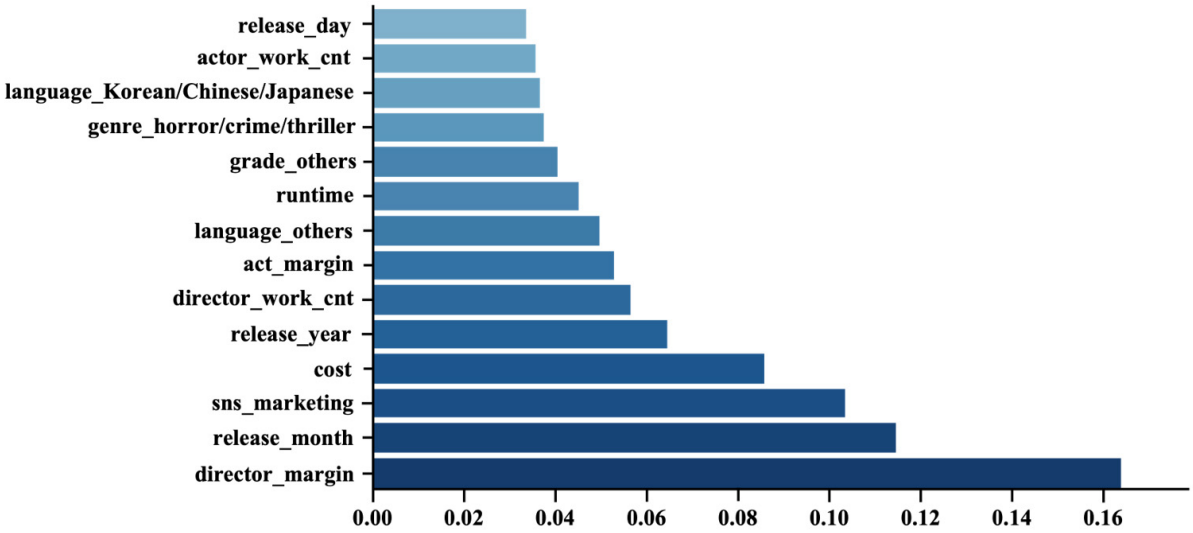

FIGURE 6. Feature importance

is successful or not. In addition, the release month, release year, cost, and online marketing variables significantly work in common on the three models.

6. **Conclusions.** This study built six models using machine learning algorithms and compared the models' performance and results. The models predict whether a movie will be a success or not from an economic point of view before its release. Comparative analysis of six models shows that the model learned with DNN has the best predictive performance. On the other hand, it takes longer to learn the model compared to other models, and it could be improved if the model is trained in a high-performance hardware environment. Based on the importance of variables extracted from three tree-based prediction models, we found that the director's reputation, the year of release, the month of release, cost, and online marketing significantly impact the movie's box office success predictions.

The variables such as the ratings, the number of screens, and reviews, have not existed and are not available at the time of the film production since those variables are generated after the movie is released. Hence, it can be challenging to predict the movie's box office success and is limited to helping managers make decisions before it is released.

This study helps the managers in decision-making from pre-release time, i.e., the planning/production stage, to the post-release stage. It also offers the advantage of reducing uncertainty in related decision-making issues by presenting predictions about whether the film will be successful or not. In future studies, more reliable and meaningful predictions are expected by adding variables that are expected to affect the film's box office success, such as scenarios and supporting actor impact, or by increasing the number of learning data per class.

Acknowledgment. Soongsil University supports this work.

REFERENCES

[1] J. A. Son and B. S. Koo, *Profitability Analysis of Korean Films in 2015*, Film Promotion Committee, 2017.

[2] D. M. L. Dissanayake and V. G. T. N. Vidanagama, Early prediction of movie success using machine learning models, *International Journal of Computer Applications*, vol.183, no.44, pp.14-21, DOI: 10.5120/ijca2021921847, 2021.

[3] Q. I. Mahmud, N. Z. Shuchi, F. M. Tawsif, A. Mohaimen and A. Tasnim, A machine learning approach to predict movie revenue based on pre-released movie metadata, *Journal of Computer Science*, vol.16, no.6, pp.749-767, DOI: 10.3844/jcssp.2020.749.767, 2020.

[4] C. Zhang, Y.-X. Tian and Z.-P. Fan, Forecasting the box offices of movies coming soon using social media analysis: A method based on improved Bass models, *Expert Systems with Applications*, vol.191, DOI: 8443/10.1016/j.eswa.2021.116241, 2021.

[5] R. Nihalaani, A. Shete and D. Khan, Movie success prediction using Naïve Bayes, logistic regression and support vector machine, *The 9th International Conference on Reliability, INFOCOM Technologies and Optimization (Trends and Future Directions)*, pp.1-5, DOI: 10.1109/ICRITO 51393.2021.9596138, 2021.

[6] M. D. Athira and K. S. Lakshmi, Movie success prediction using ensemble classifier, *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pp.1-5, DOI: 10.1109/ICCC I48352.2020.9104183, 2020.

[7] W. Lin, X. Wu, L. Lin, A. Wen and J. Li, An ensemble random forest algorithm for insurance big data analysis, *IEEE Access*, vol.5, pp.16569-16575, DOI: 10.1109/ACCESS.2017.2738069, 2017.

[8] V. Vapnik, S. E. Golowich and A. Smola, Support vector method for function approximation, regression estimation, and signal processing, *Advances in Neural Information Processing Systems*, pp.281-287, 1996.

[9] R. Mitchell and E. Frank, Accelerating the XGBoost algorithm using GPU computing, *Peer J. Computer Science*, DOI: 10.7717/peerj-cs.127, 2017.

[10] *Ensemble Method*, https://scikitlearn.org/stable/modules/ensemble.html, Accessed on May 8, 2021.

[11] N. Buduma and N. Locascio, *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms*, O'Reilly Media, Inc., 2017.

[12] F. S. Panchal and M. Panchal, Review on methods of selecting number of hidden nodes in artificial neural network, *International J. of Computer Science and Mobile Computing*, vol.3, no.11, pp.455-464, 2014.

[13] D. Anguita, A. Ghio, S. Ridella and D. Sterpi, K-fold cross validation for error rate estimate in support vector machines, *Proc. of the 2009 International Conference on Data Mining (DMIN2009)*, Las Vegas, USA, pp.13-16, 2009.

[14] N. Quader, M. O. Gani, D. Chaki and M. H. Ali, A machine learning approach to predict movie box-office success, *The 20th International Conference of Computer and Information Technology (ICCIT)*, Dhaka, pp.1-7, DOI: 10.1109/ICCITECHN.2017.8281839, 2017.