Predicting movie success with machine learning techniques: ways to improve accuracy

Lee, K., Park, J., Kim, I., Choi, Y.

Original citation & hyperlink:
Lee, K, Park, J, Kim, I & Choi, Y 2016, 'Predicting movie success with machine learning techniques: ways to improve accuracy' Information Systems Frontiers, vol (in press).
DOI: 10.1007/s10796-016-9689-z
https://dx.doi.org/10.1007/s10796-016-9689-z

DOI 10.1007/s10796-016-9689-z
ISSN 1387-3326
ESSN 1572-9419

Publisher: Springer

The final publication is available at Springer via http://dx.doi.org/[10.1007/s10796-016-9689-z

Predicting Movie Success with Machine Learning Techniques: Ways to Improve Accuracy

1. Kyuhan Lee:

Graduate School of Business, Seoul National University,

LG hall 113, Gwanak-ro 1, Gwanak-gu, Seoul, 08826, Republic of Korea

2. Jinsoo Park (corresponding)

Graduate School of Business, Seoul National University,

LG hall 113, Gwanak-ro 1, Gwanak-gu, Seoul, 08826, Republic of Korea

jinsoo@snu.ac.kr, +82028809385

3. Iljoo Kim

Erivan K. Haub School of Business, Saint Joseph's University

347 Mandeville Hall, 5600 City Ave., Philadelphia, PA 19131, USA

4. Youngseok Choi

Brunel Business School, Brunel University
Brunel Business School, Brunel University, London, UB8 3PH

Abstract

Previous studies on predicting the box-office performance of a movie using machine learning techniques have shown practical levels of predictive accuracy. Their works are technically- and methodologically-oriented, investigating what algorithms are better at predicting the movie performance. However, the accuracy of prediction model can also be elevated by taking other perspectives. For example, it is possible to increase the model accuracy by introducing unexplored features that might be related to the prediction of the outcomes. In this paper, we examine multiple approaches to improve the performance of the prediction model. First, we add a new feature derived from the theory of transmedia storytelling. Such theory-driven feature selection not only increases the forecast accuracy, but also enhances the interpretability of a prediction model. Second, we use an ensemble approach, which has rarely been adopted in the research on predicting box-office performance. As a result, the proposed model, Cinema Ensemble Model (CEM), outperforms the prediction models from the past studies using machine learning algorithms. We suggest that CEM can be extensively used for industrial experts as a powerful tool for improving decision-making process.

Keywords: Prediction model, Movie performance, Machine learning techniques, Cinema Ensemble Model, Transmedia storytelling, Feature selection

1. Introduction

The expansion of the movie industry has been a worldwide phenomenon. According to the annual report from Motion Picture Association of America, the global box-office market reached $36.4 billion in 2014. Reflecting on its economic impact, many researchers have conducted studies on the movie industry. Recently, a new research stream has emerged on box-office prediction models using machine learning techniques (e.g. Sharda and Delen 2006; Zhang, Luo, and Yang 2009; Du, Xu, and Huang 2014). The predictive nature of these studies has a significant impact on the movie industry (Simonoff and Sparrow 2000), since it provides directional guidelines to the movie producers who bear the risk of uncertainty when deciding which movies to produce. Indeed, we can cite numerous cases of failure regarding the predictions of movie success. For example, the number of audience attracted by *Mr.go*, a Korean movie produced in 2013 with the record-breaking production cost, was far below investors' expectation. The money invested in the production of *Mr.go* was about 20 million US-Dollars, and the movie was expected to attract at least five million movie-goers in Korea. However, the total attendance was less than 1.5 million according to the Korean Film Council. Thus, building a highly accurate model for predicting movie's success is a requisite to industrial decision makers who desperately wish to decrease the possibility of making false decision in green-lighting process, the process of formally approving the production of a movie.

In this study, we suggest such model that can attenuate the uncertainty in forecasting the performance of a movie. The aforementioned stream of research, which builds a prediction model for movie's success based on machine learning techniques, presents fairly high-level of prediction accuracy. However, their efforts to improve the models' prediction power have been limited only to the modification of the algorithms rather than finding meaningful features that might be critical to expect the success of movie. To elaborate, the researchers in the past have mainly focused on introducing new machine learning algorithms and testing their performances, and it was pretty much the sole objective of their studies. Although such efforts have contributed to the increase of the prediction accuracy, we believe that the accuracy can be further increased by taking other perspectives. For example, it is possible to introduce an unexplored feature to a prediction model or to implement a feature-selection for existing features.

Generally, the feature selection is one of the frequently-considered methods to increase the performance and the interpretability of machine learning algorithms. However, in this study, we focus more on introducing a new

feature rather than pruning the expectation model with existing ones. The reasoning behind our decision is that the features used in our study already have been tested to be highly effective for predicting a movie success in the past research. Thus, we expect that the exclusion of some of such features will decrease the accuracy of the prediction model. In addition, we have considered that the number of features used in this study is not as many to the extent that it deteriorates the performance of a prediction model. For example, in studies, as ones in biology, the use of more than thousands of features drastically decreases the model accuracy and interpretability, and requires exponentially-increased model training and testing time (Guyon and Elisseeff 2003). However, since we include only twenty-one features derived from six types of variables, we have considered that it is unnecessary to remove a part of features in study.

Thus, we, rather than eliminating the abundant features, introduce an unexplored feature that may increase the accuracy of our prediction model. To elaborate, we investigate the impact of a new feature, based on the theory of transmedia storytelling, on the outcome. This is the first study to include the transmedia storytelling as a feature for the movie success prediction. According to our experiment result, the introduction of transmedia storytelling feature has boosted the performance of our prediction model. Also, the introduction of the new feature based on a solid theoretical background will allow us not only to elevate the accuracy of prediction model but also to increase the explanatory power of the model. By selecting the feature based on such theory, we can better justify and explain the causal relationship between the feature and the outcome.

In addition to the aforementioned feature-oriented approach, we, in this study, also consider methodology-driven approach to improve the prediction accuracy. In detail, we use an ensemble approach to build a better-performing prediction model. The effect of the ensemble approach in enhancing the model accuracy has been widely recognized in academia (Elder 2003). However, few, if any, studies have used the ensemble method in building a prediction model for movie's success.

The rest of this article is organized as follows. Section 2 provides critical reviews on the past research on predicting movie's success. In Section 3, we suggest the detailed descriptions on the methodology implemented in this study. In the following section (Section 4), the information on the data used in this study is given. Then, in Section 5, we suggest the results of the prediction model built in this study. Finally, research implications and future research are discussed in Section 6.

2. Related Works

2.1. Predictive studies in the movie domain

Most of the past studies regarding the movie industry have had the explanatory nature, investigating factors that affect the box-office performances of movies. The earliest works include the research conducted by Litman (1983). He has investigated how the production cost, critics' ratings, genre, distributor, release season, and main actor's award history are related to a movie's box-office performance. As the movie industry has kept growing since the Litman's study, the exploration of factors affecting movie success has been an interesting research area and thus abounding articles have been published within the area. Vany and Walls (1999), Elberse (2007), and Nelson and Glotfelty (2012) have examine the relationship between a main actor's star power and a movie performance. Basuroy, Chatterjee, and Ravid (2003) have investigated how critical reviews affect a movie success, setting star power and budgets as moderators. Prag and Casavant (1994) have had an interest in identifying the relationship between factors such as marketing costs, MPAA ratings, and sequels and a movie success.

Recently, based on the knowledge accumulated from these studies, a few researchers have begun to conduct the studies that have the predictive characteristic. For example, forecasting the movies that are highly possible to succeed is one of the types of such research. Asur and Huberman (2010) have used Twitter data to predict a movie success and Mishine and Glance (2006) have predicted movie sales using web blog data. Especially, using machine learning techniques, several studies have produced the prediction models with the moderate level of accuracy (e.g. Sharda and Delen 2006; Eliashberg, Hui, and Zhang 2007; Zhang, Luo, and Yang 2009; Du, Xu, and Huang 2014). For instance, Sharda and Delen (2006) have examined the performance of the logistic regression, discriminant analysis, decision tree, and neural networks to forecast movie's success. They have used MPAA ratings, competition level, main actor's star value, genre, special effects, sequel, and the number of screens at the initial day of movie release as features to predict the movie performance. Their best-performing model has predicted the nine outcome variables with the 36.9% of accuracy. Zhang, Luo, and Yang (2009) have suggested a multi-layer back propagation neural network that has improved the neural network model presented by Sharda and Delen (2006). Their model correctly has classified six outcome variables with 47.9% of accuracy. Eliashberg, Hui, and Zhang (2007) have forecasted a movie's return on investment based solely on its script information using the decision tree algorithm. Du, Xu, and Huang (2014) have evaluated the performance of the linear regression, support vector machine, and neural networks on predicting

the box-office success, analyzing the sentiments of the texts posted on Tencent Microblog. The summary of the representative previous research in the movie domain is shown in Table 1.

Table 1 Summary of Previous Research

	Author(s)	Features Considered	Methods Used
Explanatory Research	Litman (1983)	Production cost, critics' rating, genre, distributor, release season, main actor's award history	Regression Analysis
	Prag and Casavant (1994)	Marketing cost, quality, star value, sequel, award, genre, MPAA reting	Regression Analysis
	Basuroy, Chatterjee, and Ravid (2003)	Critical review, star power, budget	Regression Analysis
	Nelson and Glotfelty (2012)	Star power	Regression Analysis
Predictive Research	Sharda and Delen (2006)	MPAA rating, competition, star value, genre, special effects, sequel, number of screens at the initial day of release	Logistic Regression, Discriminant Analysis, Classification and Regression Tree, Neural Networks
	Eliashberg, Hui, and Zhang (2007)	Movie script	Classification and Regression Tree
	Zhang, Luo, and Yang (2009)	Nation, director, performer, propaganda, content catregory, month, week, festival, competition, cinema number, screen number	Neural Networks
	Du, Xu, and Huang (2014)	Microblog posting counts, microblog posting content	Support Vector Machine, Neural Networks

While these studies have mostly focused on methodological perspective to improve their model accuracy, we suggest more comprehensive method that enhances the performance of the model. In this study, we implement

both the feature-oriented and methodology-driven approach. First, we introduce a new feature derived from the solid theory of transmedia storytelling. Second, we use an ensemble learning method that has hardly been applied to the research in the movie domain. In the following sections, we provide a detailed explanation on the theory of transmedia storytelling as well as the process of constructing the ensemble model.

2.2. The theory of transmedia storytelling

Transmedia storytelling refers to the delivery of a single story across multiple media channels such as television, books, and games. The contents on different channels provide distinctive and independent experiences, but essentially people consume them in a coordinated way (Edwards 2012). If such contents interact with each other and evolve to be a transmedia story, it may produce a synergy effect, forming a richer background story and attracting a wider audience (Jenkins 2003). This transmedia storytelling is "one of the most important sources of complexity in contemporary popular culture" (Scolari 2009, p 587). Transmedia storytelling improves the consumer experience of not only the content it carries but also the content that other media transfers.

The theory of transmedia storytelling is not aa new concept and has been adopted both in industries and academia. For instance, in the entertainment industry, horizontally-integrated media companies, such as Warner Brothers that owns DC Comics, possess multiple channels that can be used to deliver a single story, and they are highly motivated to brand their products through as many channels as possible (Jenkins 2007). In academia, since Jenkins (2003) has first suggested the term 'transmedia storytelling' to refer to a complete story delivered through multiple but connected media (Blumenthal and Xu 2012), much research has been conducted regarding the concept. Long (2007) and Perryman (2008), through the case studies, have identified how the transmedia storytelling is deployed in the real world. Blumenthal and Xu (2012) have investigated the four components needed to be considered when designing a transmedia story. Moloney (2011) has examined the possibility of adopting the transmedia storytelling strategy in a journalism context. He expects that journalists can better engage publics through adopting the strategy.

Although the research regarding the transmedia products has been conducted more than a decade, there is not much consensus on the characteristics of transmedia stories. However, we consider that Dena (2004) provides precise explanation on such characteristics. She suggests that transmedia works should possess the following features: (1) user activity, (2) narrative-driven activity, and (3) navigation between media. To elaborate, first, the consumer of

a transmedia work has to show an effort to assemble the scattered information on the story across multiple media. For example, one who has seen the movie Iron Man may be willing to seek further information on the story of the Iron Man through other media such as comic books. Second, this consumer participation should be directed by the story itself. That is, the consumer participates because each medium that delivers the story of Iron Man refers to one another to form a complete story. Although the story delivered via each medium makes sense by itself, it also provides a piece of information to understand the bigger story. Third, the consumer's navigation between media can be classified into the following two types: (1) navigation across different channels and (2) navigation across different modes within a channel. The channel here is a concept combining a medium and its environment. For example, a standard movie theater and an IMAX movie theater delivers a story through the same medium, film, but under the different environment. Then, the mode refers to the way that a story is delivered. For example, an audio file and a video file possess different modalities. The user can experience different modes within a single channel. For instance, the user can read people's complementary comments on the Iron Man on the movie review website, and watch the movie trailer on YouTube. In this study, we have tried to identify transmedia works that satisfy Dena's definition. However, the criteria regarding the user activity and the narrative-driven activity are hard to identify unless we closely analyze each movie's content. Thus, in this study, we have only adopted the navigation between media as the only criterion to classify movies based on the transmedia storytelling strategy.

3. Methodology

3.1. Building an Ensemble Model for Predicting Movie Success

According to Dietterich (1997), there are several classic approaches to construct an ensemble model. First, we can subsample training sets, build different classifiers on each set, and combine the estimates of these classifiers. Second, we may use different subset of features to make different classifiers and combine their estimates. Third, it is possible to manipulate the output targets to build multiple classifiers and merge them into one.

In this study, we use a different approach to build an ensemble model. To elaborate, we first build candidate classifiers for the ensemble model using seven different algorithms. The rationale for inclusion of these algorithms is suggested in the subsequent section. Among the candidates, we select ones that present relatively high level of prediction accuracy. Then, we build an ensemble model by voting the estimates of each component model. In this

paper, we use a plurality voting system in which the winning estimate is the one that with the largest votes. Through such process, the Cinema Ensemble Model (CEM), an ensemble model for the prediction of movie's success, can be constructed. The process is schematized in Fig. 1.

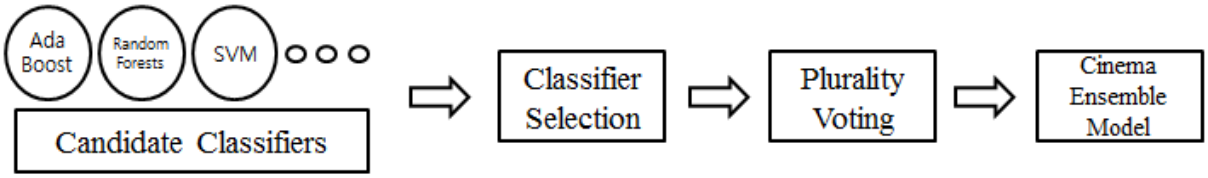

Fig. 1 The Process of Building CEM

It is also important to note that some of the candidate classifiers in this study are themselves ensemble models. For example, Ada Tree Boosting, Gradient Tree Boosting, and Random Forests are ensemble algorithms. Thus, CEM is an ensemble model constructed upon other ensemble methods. It can be considered to be the 'ensemble-of-ensemble.'

3.2. Descriptions of Learning Algorithms for Component Models

As explained above, seven machine learning algorithms are used to build candidate models: *Adaptive tree boosting, gradient tree boosting, linear discriminant, logistic regression, neural networks, random forests, and support vector classifier.* We have carefully and comprehensively reviewed previous research applying machine learning techniques on the classification problem, and selected these seven algorithms. Especially, unlike other existing research pertaining to the movie domain, our research has utilized the most types of algorithms for the comparison of performance. In other words, we have considered, to the best of our knowledge, all the classification algorithms that have been used in the past research suggesting prediction models for a movie performance. The brief description of the algorithms used here is presented in the following.

3.2.1. Adaptive Tree Boosting

Adaptive tree boosting (ATB) is the algorithm of which the concept is based on *boosting*. *Boosting* is a method to improve the performance of an algorithm by producing multiple classifiers and combining the estimates of these classifiers (Freund and Schapire 1999). Although each classifier is moderately inaccurate, the model accuracy is high when combined altogether. In such fashion, *ATB* produces a number of weak classifiers whose error rate is slightly better than random guessing. Each classifier is consecutively built after one another using a modified set of training data. To specify, if we suppose *ATB* builds the weak classifiers for *t* rounds, at each round, the weights of data points

are adjusted based on whether the points are correctly classified in the previous round. For the points that are incorrectly classified, the weights are increased so that the weak classifier can be trained focusing on such points (Hastie 2005; Freund and Schapire 1999). The performance of *ATB* algorithm has been widely recognized, and especially it is well-fitted to multi-class classification problems (Zhu et al. 2009). Thus, we include *ATB* as one of the algorithms to build candidate models.

3.2.2. Gradient Tree Boosting

Gradient tree boosting (GTB) works in a similar way to *ATB* in that it builds, at each round, a classifier using residuals of the previous prediction function (Yamagashi, Kawai, and Kobayashi 2008). However, *GTB* differs from *ATB* that it uses a different measure (i.e., binomial deviance) to determine the cost of errors (Hastie et al. 2009; Chamber and Dinsmore 2014). It is commonly accepted that *GTB* is robust with the problem in which a multicollinearity issue exists and the number of features is relatively large to the number of data points (Mayr et al. 2014; Prettenhofer and Louppe 2014). Since, in this study, we have collected 375 data points with 21 variables (i.e., 21 variables derived from 6 features), we assume that *GTB* can produce reliable results with our data set.

3.2.3. Linear Discriminant

Linear discriminant (LD) is one of the commonly used algorithms for data classification. *LD* extracts the classification criterion from data sets (Zhang 2003). By this criterion, the between class variance is maximized while the within class variance is minimized (Balakrishnama and Ganapathiraju 1998). If the assumption of normality for the data is fulfilled, *LD* produces robust and reliable results even when the sample size is small. In addition, the robustness of *LD* remains with the multiple target variables (Pohar, Blas, and Turk 2004). Thus, we consider *LD* as one of the candidate algorithms that may be suitable to our multi-classification problem.

3.2.4. Logistic Regression

Logistic regression (LR) is one of the most widely-used algorithms to predict binary outcomes. The prediction is based on the probability calculated by the logistic function which ranges between 0 and 1. Although *LR* is commonly used to explain the relationship between multiple predictor variables and dichotomous dependent variables, it can also be applied to the problems with multi-categorical dependent variables (Kleinbaum and Klein 2010). There exist several methods, such as one-vs-all and one-vs-one strategy, to convert a binary classification problem into a multiple classification problem. In this study, we use one-vs-all strategy, which fits one classifier per class against all

the other classes (DeMaris 1995). Unlike *LD*, *LR* makes no assumption regarding the normal distribution of sample data. Thus, it is more flexible and robust with the data that do not fulfil the normality assumption (Pohar, Blas, and Turk 2004).

3.2.5. Neural Networks

Artificial Neural networks (ANN) is a machine learning technique receiving much public attention recently. Since *ANN* typically requires longer training time and its learned target function is hard to interpret (Mitchell 1997), it has not been a popular method comparing to others such as decision tree. However, with the exponential growth of the computing power and the algorithm's strong performance, nowadays *ANN* and its variations have been widely used both in academia and the industry. In this study, we use multilayer perceptron (*MLP*) with four layers including input layer, output layer, and two hidden layers. It is widely accepted that *MLP* can effectively express nonlinear decision surfaces (Mitchell 1997).

3.2.6. Random Forests

Random forests (RF) is an algorithm that makes a prediction by combining the estimates of randomly-built independent decision trees (Breiman 2001). Although it has less interpretability than an individual tree, it is widely recognized that *RF* presents significantly better performance. At the same time, *RF* is robust to outliers and has a good ability to deal with irrelevant inputs (Montillo 2009). We expect *RF* can produce a candidate model with high prediction accuracy.

3.2.7. Support Vector Classifier

Support Vector Classifier (SVC) aims to find the maximum-margin hyperplanes that optimally separate the classes in the training data (Auria and Moro 2008). *SVC* has the advantages that it shows strong generalization ability and is robust to outliers (Abe 2005). It is one of the most widely used machine learning algorithm these days. It is utilized to improve the performance of the medical diagnostics, optical character recognition, and many other fields.

3.3. Discretization of the Movie Success

In this study, we define the prediction of box-office success as a classification problem. This strategy has been applied in a few past studies (e.g., Sharda and Delen 2006; Zhang, Luo, and Yang 2009). We discretize the dependent variable (i.e. box-office performance) into six classes. The range for each class is determined based on the interviews with industry experts. Since a budget for each movie is different, we cannot generalize a break-even-point (BEP) of the

movie. According to the experts, BEP attendance commonly exists within the range of class 3. However, for the movies with large amount of investment, their BEPs can be within the range of upper classes. The breakpoints used to discretize the dependent variable are shown in Table 2.

Table 2 Movie Performance Classes

Class	Attendance Range (in thousands)	Revenue Range (Approx. in $ thousands)
1(Blockbuster)	> 4,000	> 26,700
2	2,000 - 4,000	13,300 - 26,700
3	750 - 2,000	5,300 - 13,300
4	250 – 750	1,800 - 5,300
5	100 – 250	700 - 1,800
6(Flop)	< 100	< 700

3.4. Feature Description

We use six different types of features in this study. We have selected the features including the ones that widely used in the past studies. In addition, the cadre of a Korean film production and distribution company has verified whether our selection of features is comprehensive enough to successfully predict a movie's performance.

We note that categorical features with more than two possible values are converted into *n*-binary features, where *n* represents the number of the values. For example, *genre*, one of the features in this study, has sixteen possible values including ACTION, ADVENTURE, COMEDY, and so on. We convert these values into sixteen-binary features so that each feature is set to either 0 or 1. To elaborate, when a movie is assigned to two categories – ACTION and COMEDY, the values of these two features are set to 1, and the values of the other fourteen features are set to 0. The following sub-sections describe the features included in this study.

3.4.1. Genre

Genre is one of the most basic and commonly used variables in predicting a movie's success (Sharda and Delen 2006). In this study, we use the sixteen categories suggested by the Korean Media Rating Board (KMRB) to classify each movie. Each movie can be classified into multiple genres. The genres included in this study are as follows: ACTION, ADVENTURE, ANIMATION, COMEDY, CRIMINAL, DOCUMENTARY, DRAMA, EPIC, FAMILY, FANTASY, HORROR, INDEPENDENT, MYSTERY, ROMANCE, SF, and THRILLER. The information on movie genres has been collected from the webpage of the KMRB.

3.4.2. Sequel

The impact of sequels on a movie's success is also well-recognized by practitioners. Movie producers often produce sequel movies to reduce risk and uncertainty (Eliashberg, Elberse, and Leenders 2006). For example, the Marvel Studios has produced a sequence of movies under the series name of *Avengers*. The series have been successful not only in the North American market but also worldwide. Besides, Dhar, Sun, and Weinberg (2012) have identified that sequels have a positive impact on both supply and demand side of movie distribution. More often than not, a sequel movie tends to be distributed to a significantly larger number of theaters (i.e., positive impact on the supply side). Also, sequels tend to attract more movie-goers than non-sequels (i.e., positive impact on the demand side). Thus, we include *sequel* as an important feature to predict a movie's success. It is necessary to note that we do not consider the movie that has been remade as a sequel of the original movie since such a movie is unlikely to be helpful in discriminating between demographic classes.

3.4.3. Number of Plays at the Initial Day of Release

Several past studies have used *the number of screens at the initial day of release* as one of the features for their prediction models (e.g., Sharda and Delen 2006; Zhang, Luo, and Yang 2009; Ghiassi, Lio, and Moon 2015). The industry experts that we interviewed also pointed out that the number of screens is an effective predictor of movie's success.

In this study, we use *the number of plays at the initial day of release*, instead of *the number of screens at the initial day of release*, as a feature for our prediction model. The rationale for our decision is that *the number of screens at the initial day of release* does not reflect the running time of a movie. This may result in the misinterpretation on the influence of the number of screens, because two movies with different running times may vary in their numbers of plays even when the numbers of screens for the both movies are exactly the same. Such different numbers of plays mean distinctive levels of exposure to movie goers, affecting movies' performances. For example, the movie *The Martian* with the running time of 144 minutes may be shown less number of times a day than the movie *The Good Dinosaur* with the running time of 100 minutes. Consequently, *The Good Dinosaur* has higher possibility to succeed in box-office if all the other factors affecting movie's performance are controlled.

Our data on *the number of plays at the initial day of release* has been collected from the webpage of KMRB. KMRB tracks and provides the information on the daily number of screens and plays of a movie for its entire screening period.

3.4.4. Movie Buzz before the Release

Movie buzz is the feature that has been recently highlighted. For example, Mishne and Glance (2006) has made a prediction of movie sales using the buzz data on web blog. Liu (2006) has identified the explanatory power of movie buzz in box-office prediction. In Liu's research, he describes the volume of buzz as the major factor that explains box-office performance. In this study, we include the number of movie comments (i.e., *movie buzz)* on *Naver Movie* (see http://movie.naver.com/) as one of the features for our prediction model. The *naver.com*, the most popular search engine site in Korea, has a movie page showing various types of information on movies. An example of the movie page is presented in Fig. 2. On the movie page, there is a review section where people can write comments before and/or after the movie release. From this section, we count the number of comments that have been written before the movie release.

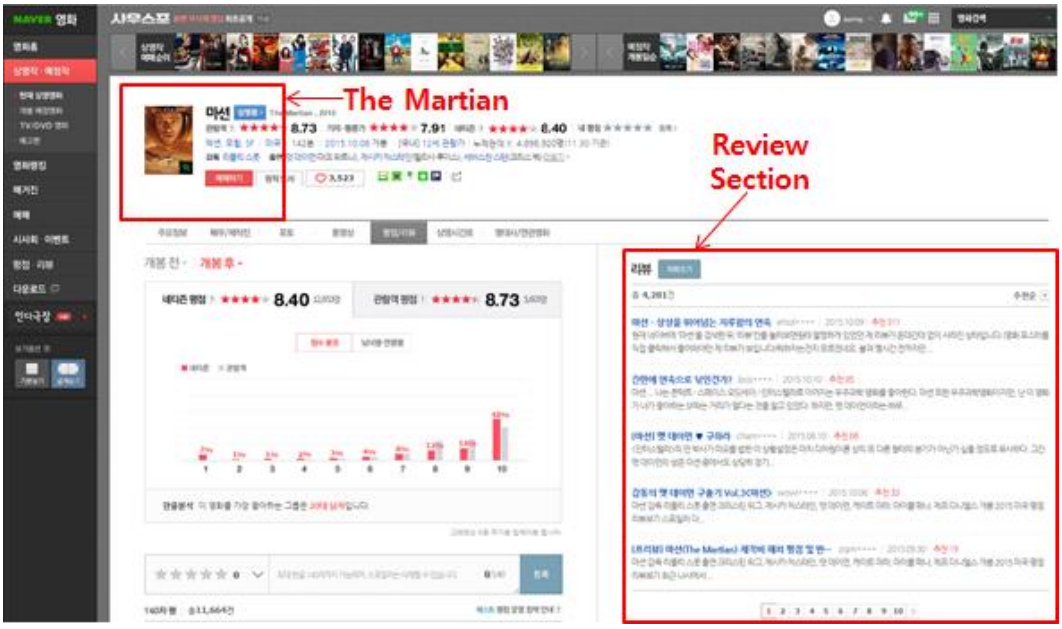

Fig. 2 *The Martian*'s Page on *Naver Movie*

3.4.5. Transmedia Storytelling

As mentioned above, we have considered the movies based on television series, novels or comics to be the ones implementing the transmedia storytelling strategy. For the foreign movies, we have used the data provided by *IMDB.com*[1]. For the domestic movies, we have used the information presented on *Naver Movie*. Either 0 or 1 is assigned as the value of *transmedia storytelling*. When the writing credit goes solely to a single or multiple screenplay

[1] IMDB.com is the most popular website that provides movie-related information in the U.S.

writer(s), 0 is assigned, and when the movie is based on the story from other media, 1 is assigned. We have not considered remade movies the ones that implement the transmedia storytelling strategy.

3.4.6. Star Buzz (i.e., Star Power)

Although a plethora of research has been conducted to identify the impact of stars on movie's success (e.g., Ravid 1999; Elberse 2007; Nelson and Glotfelty 2012; Treme and Craig. 2013), the empirical findings of their research show mixed results. There may be multiple reasons for such inconsistent results, but the most explicit cause is the use of different metrics for measuring the star power. For example, while Academy Award wins and nominations have been widely used as a proxy for the star power (e.g. Litman 1983; Ravid 1999; Basuroy, Chatterjee, and Ravid 2006), there are other metrics that are alternatively utilized to measure star power. Nelson and Glotfelty (2012) have used STARmeter rankings from *IMDB.com*. Treme and Craig (2013) have used the number of times that actors/actresses appear in *People* magazine before the movie release.

In addition, each of these metrics involves limitations. First, Academy Award wins and nominations highly limit the number of actors/actresses who are classified as stars (Nelson and Glotfelty 2012). Second, since the STARmeter rankings change weekly, it only gives fragmented information on star power at a point, making it hard to track star power spanning more than a week. Lastly, stars' appearance on *People*, as Academy Award wins and nominations, limits the scope of actors/actresses whose star power can be empirically measured.

In this study, we use online *star buzz* as an appropriate proxy to measure the star power. We have counted the number of posts on *Naver Blog*[2] in which stars are referred. We find this metric compelling since it does not reveal any of the weaknesses mentioned above. In other words, it can measure the star power with infinite number of actors/actresses over any period of time.

Since movie producers and distributors generally start to promote movies a month before their release, it will be advantageous for them to know the expected performance of the movies in advance to the outset of the promotion. Thus, we have collected *star buzz* data from two months before the movie release to a month before the movie release.

[2] *Naver Blog* is the most popular personal blog site in Korea. Individuals mostly use it as a way to express their thoughts and communicate with others. Commercial companies also utilized *Naver Blog* with the purpose of advertising their products and services. (see http://blog.naver.com)

4. Data Collection

The data used in our study includes movies that are released from October 25, 2012, to December 31, 2014. The data has been collected from the Korean Film Council webpage and *naver.com*. We have considered only the top 400 movies by the number of viewers, because including movies beyond the top 400 can lead to a 'spurious improvement' of the prediction models. That is, since all movies beyond the top 400 are categorized into the same class (i.e. 'flop' class; refer to Table 2), the inclusion of those movies tends to increase the probability of correct classification. Furthermore, through the interview with decision makers from a film production company, we have found that practitioners are far more interested in predicting the performance of 'major' movies whose budgets are usually more than two million US dollars. The performances of these movies do not usually fall into the 'flop' class even in the worst cases. Thus, we assume that including movies beyond the top 400 is unnecessary. Among the 400 movies, excluding movies that have missing values leaves us with 375 movies. A summary of the statistics from the collected data is presented in Fig. 3.

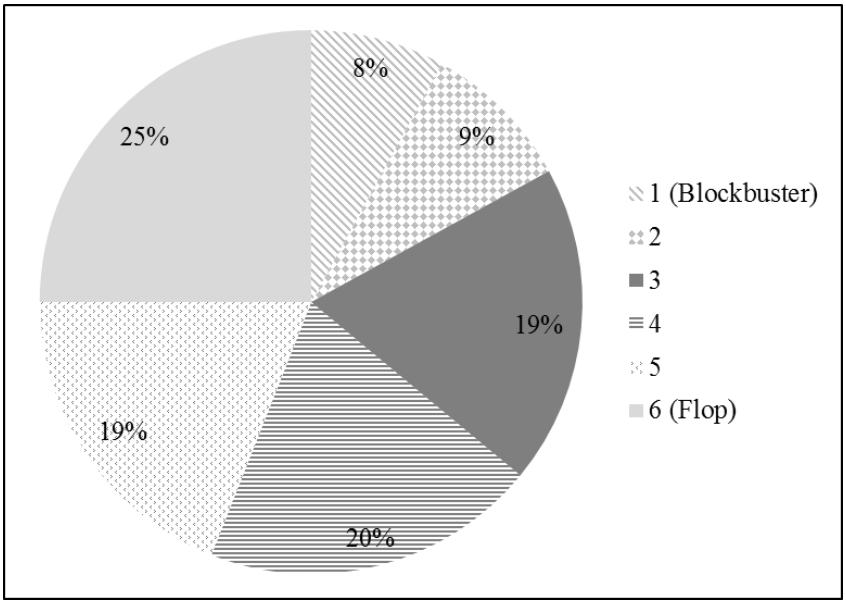

Fig. 3 Distribution of Movie Classes

5. Analysis

5.1. Performance Metrics

In this study, we adopt the performance metrics of Sharda and Delen (2006). They have used Average Percent Hit Rate (APHR) to measure the accuracy of their prediction models. Two different types of APHRs are calculated in this

study: Bingo and 1-Away. Bingo counts the number of classifications that exactly matches their actual classes, 1-Away represents within-one-class hit rate. For example, if CEM predicts a movie to be in the class 1 and the actual outcome of the movie belongs to the class 1, it is classified as Bingo. On the other hand, if CEM predicts the movie to be in the class 2 and the actual outcome of the movie belongs to the class 1 or 3, the prediction is missed by one class so that it is classified as 1-Away. If a prediction is missed by more than one class, we consider it to be a misprediction. Two APHRs can be formulated in the following equations:

$$\text{APHR} = \frac{Number\ of\ test\ \square\square\square\square\ points\ correctly\ classified}{Total\ number\ of\ test\ \square\square\square\square\ points},$$

$$APHR_{Bingo} = \frac{1}{n}\sum_{i=1}^{g} p_i,$$

$$APHR_{1-Away} = \frac{1}{n}(\sum_{i=1}^{g}(p_{i-1} + p_i + p_{i+1}) - (p_0 + p_{g+1})),$$

where g is the total number of classes (i.e. $g = 6$), n is the total number of test data points (i.e. $1 \leq n \leq 375$), and p_i is the total number of data points correctly classified as class i. In the case of $APHR_{1-Away}$, we define $(p_{i-1} + p_i + p_{i+1})$ as the total number of data points correctly classified as class i. These metrics have been used not only in Sharda and Delen's research but also in Zhang, Luo, and Yang (2009). By using the same metrics as the ones used in the past two studies, we are able to compare our model to the previous ones and identify whether our approaches have improved the model performance.

5.2. Candidate-model Performance

As mentioned above, we build seven candidate models based on different machine learning algorithms. The performance of each model has been evaluated by repeated random sub-sampling validation method. This method repeats the validation with the random partitions of training data and test data. Repeated random sub-sampling validation resolves the issue of *k*-fold cross validation that the size of test data shrinks as *k* grows, increasing the performance variance of each individual fold (Thornton et al. 2012). The influence of such issue can deteriorate when the volume of data is small. Since the size of the data set in this study is limited, we have concluded that repeated random sub-sampling validation is far more suitable than *k*-fold cross validation. We have repeated the validation process ten times with an 80/20 split of training and test dataset.

Table 3 ranks the candidate models based on two metrics: Bingo and 1-Away. The detailed result of model performance is shown in Table 4. According to the result, we find that *GTB* has performed the best for APHR Bingo. *GTB* has correctly classified 55.1% of the movies from the test dataset. *RF* has shown the second highest APHR Bingo. It has correctly classified 53.1% of the movies. *LR* and *LD* have presented moderate levels of APHR Bingo, 49.7% and 48.5% respectively. *NN* and *ATB* have not performed well in this movie prediction problem. *NN* has predicted the movie performance with 42.4% of accuracy, and *ATB* has shown 40.8% of APHR Bingo,

Table 3 Model Performance Rank

Rank	Bingo	1-Away
1	Gradient Tree Boosting	Gradient Tree Boosting
2	Random Forests	Logistic Regression
3	Logistic Regression	Adaptive Tree Boosting
4	Linear Discriminant	Random Forests
5	Neural Networks (Multilayer Perceptron)	Linear Discriminant
6	Adaptive Tree Boosting	Neural Networks (Multilayer Perceptron)
7	Support Vector Classifier	Support Vector Classifier

Table 4 APHRs of Six Candidate Models

	ATB		***GTB***		***LD***		***LR***		***NN (MLP)***		***RF***		***SVC***	
Rep.	Bingo	1-Away	Bingo	1-Away	Bingo	1-Away	Bingo	1-Away	Bingo	1-Away	Bingo	1-Away	Bingo	1-Away
1	37.3%	89.3%	58.7%	85.3%	53.3%	85.3%	36.0%	85.3%	48.0%	86.7%	46.7%	84.0%	26.7%	64.0%
2	46.7%	92.0%	46.7%	92.0%	41.3%	88.0%	45.3%	93.3%	40.0%	88.0%	56.0%	90.7%	30.7%	62.7%
3	33.3%	78.7%	56.0%	86.7%	53.3%	88.0%	61.3%	88.0%	38.7%	84.0%	53.3%	90.7%	26.7%	56.0%
4	40.0%	88.0%	52.0%	89.3%	50.7%	85.3%	54.7%	86.7%	50.7%	88.0%	56.0%	86.7%	26.7%	48.0%
5	42.7%	92.0%	57.3%	93.3%	54.7%	89.3%	56.0%	90.7%	42.7%	81.3%	62.7%	89.3%	26.7%	61.3%
6	50.7%	88.0%	54.7%	88.0%	56.0%	80.0%	54.7%	89.3%	36.0%	82.7%	49.3%	88.0%	41.3%	74.7%
7	40.0%	85.3%	57.3%	86.7%	53.3%	88.0%	50.7%	90.7%	49.3%	85.3%	57.3%	81.3%	29.3%	44.0%
8	41.3%	84.0%	56.0%	85.3%	41.3%	82.7%	45.3%	86.7%	34.7%	75.7%	49.3%	86.7%	28.0%	65.3%
9	38.7%	85.3%	57.3%	90.7%	34.7%	86.7%	42.7%	89.3%	45.3%	86.7%	50.7%	90.7%	25.3%	53.3%
10	37.3%	84.0%	54.7%	85.3%	46.7%	80.0%	50.7%	82.7%	38.7%	81.3%	49.3%	76.0%	25.3%	60.0%
AVG	**40.8%**	**86.7%**	**55.1%**	**88.3%**	**48.5%**	**85.3%**	**49.7%**	**88.3%**	**42.4%**	**84.0%**	**53.1%**	**86.4%**	**28.7%**	**58.9%**
SD	5.0%	4.1%	3.5%	2.9%	7.2%	3.4%	7.5%	3.1%	5.7%	3.9%	4.9%	4.8%	4.8%	8.9%

In addition, *GTB* and *LR* have performed the best for APHR 1-Away. 88.3% of the movies are classified correctly or misclassified by one class (i.e., 1-Away) by these algorithms. Models by *AT`B, RF, LD,* and *NN* have shown moderate levels of accuracy, reporting 86.7%, 86.4%, 85.3%, and 84.0% of APHR 1-Away, respectively. In both metrics, *SVC* has not performed well, reporting 28.7% of APHR Bingo and 58.9% of APHR 1-Away.

5.3. Cinema Ensemble Model (CEM) Performance

As an effort to improve the accuracy of predictions, we introduce CEM. As noticed earlier, we first select the appropriate candidates as the component models for CEM. According to the result in the previous section, *GTB*, *LD*, *LR*, and RF have shown good performance in predicting a movie's success. Thus, we include these four models as component models.

Each of the component models produces its own estimates (i.e., predicted classes of movies). To build CEM, we combine these estimates. In ensemble approach, the combination of estimates can be done by various strategies including voting and averaging (Elder 2003). In this paper, we use plurality voting system in which the winning estimate is the one that gets the largest votes. When two or more classes have the same number of votes (e.g., two votes for blockbuster and two votes for flop), we choose the class which *GTB* votes. Such criterion is plausible since *GTB* is one with the highest accuracy among the candidate models. To validate the result, we have applied the repeated random sub-sampling validation method. The result is shown in Table 5.

Table 5 APHRs of CEM

CEM		
Rep.	Bingo	1-Away
1	60.0%	88.0%
2	56.0%	88.0%
3	61.3%	92.0%
4	62.7%	92.0%
5	48.0%	88.0%
6	58.7%	82.7%
7	56.0%	94.7%
8	58.7%	88.0%
9	62.7%	82.7%
10	61.3%	86.7%
Mean	**58.5%**	**88.3%**
SD.	4.4%	3.9%

When compared to the performances of component models, CEM improves APHR Bingo of *GTB*, the best performing component model, by 3.4%. However, APHR 1-Away has not shown significant improvement in the

ensemble model. Comparing to the performances of the models from past studies, our model also presents enhanced result. In the study of Sharda and Delen (2006), the best performing model has showed 36.9% of APHR Bingo and 75.2% of APHR 1-Away. Our model improves the APHR Bingo by 21.6% and APHR 1-Away by 13.1%. Another study by Zhang, Luo, and Yang (2009) suggests that their model predicts the movie success with 47.9% of APHR Bingo and 82.9% of APHR 1-Away. Our model increases the accuracy of their model by 10.6% in APHR Bingo and 5.4% in APHR 1-Away.

5.4. Performance Improvement by Transmedia Storytelling Feature

The models from the previous section use all the features including *transmedia storytelling* to make a prediction. In this section, to investigate the impact of *transmedia storytelling* on model performance, we exclude *transmedia storytelling* feature from our data sets. Then, we train a CEM model with the data and examine its performance with test data. The performance of such model, CEM without *transmedia storytelling*, is shown in Table 6.

Table 6 APHRs of CEM without *Transmedia Storytelling*

CEM w/o TS		
Rep.	Bingo	1-Away
1	62.7%	90.7%
2	58.7%	96.0%
3	49.3%	88.0%
4	64.0%	86.7%
5	53.3%	88.0%
6	49.3%	90.7%
7	57.3%	88.0%
8	50.7%	86.7%
9	42.7%	84.0%
10	49.3%	93.3%
Mean	**53.7%**	**89.2%**
SD.	6.8%	3.5%

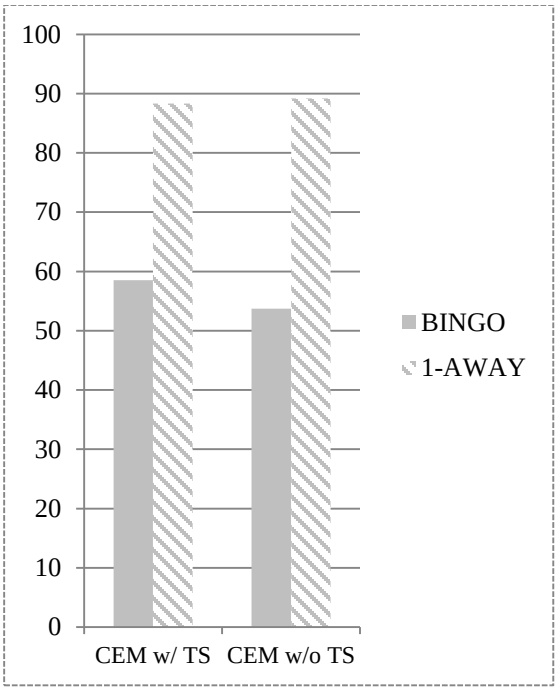

Fig. 4 Model Performance Comparison

As depicted in Fig. 4, we find that *transmedia storytelling* increases APHR Bingo of CEM by 4.8%. However, APHR 1-Away is decreased by 0.9%. Since APHR Bingo is the primary criterion for evaluating the performance of a prediction model and the decrease of the APHR 1-Away in our CEM is not significant, we conclude that *transmedia storytelling* increases the accuracy of the prediction models in this study. In addition, considering the fact that most of the features in a machine learning classifier are responsible for only the fraction of the classification performance, we argue that approximately 5% of the increase in accuracy is significant. For example, Adamopoulos (2013) describes that seven out of the eight features included in his classification model for predicting a student's online course completion have contributed less than 3% of the total accuracy respectively.

6. Discussion and Conclusion

This research presents a model for predicting box-office performances of movies. Cinema Ensemble Model (CEM) is proposed for the improvement of prediction accuracy. In addition, a new feature, *transmedia storytelling*, is introduced based on its solid theoretical background. As a result, our model has forecasted movie's success with the accuracy of 58.5%, enhancing the performances of the models from past studies.

Our study has several good implications both academically and practically. First, to the best of our knowledge, our research, among the studies forecasting a movie success with machine learning techniques, is one of the few studies that have focused on the feature aspect of a prediction model. Especially, we suggest an idea of choosing features based on concrete theories. Such theory-driven feature selection is especially compelling in that, unlike explanatory studies, most predictive studies using machine learning techniques tends to focus only on the enhancement of predictive power. In other words, they emphasize more on the construction of better-performing model, not paying much attention to the explanation of how the model's features are related to its outcome. This causes the blame on the black-box nature of machine learning techniques. However, by determining what features to include based on concrete theories, we can defend such negative critiques. Second, we identify which machine learning algorithms are suitable to movie domain and build a prediction model, CEM, based on the ensemble approach which has rarely been adopted in the previous studies. CEM has increased the prediction accuracies of past studies by at least 10%.

Our study also has a good practical implication for the decision makers in movie industry. For movie producers, our model can be used as a supplementary tool for green-lighting processes. For distributors and theater owners, the model can provide an effective way to determine which movie to select, distribute, promote, and play.

In the future work, we plan to implement a few strategies to enhance our model further. First, a more sophisticated voting criterion can be used for building an ensemble model. For example, weighted-voting criterion can be considered to increase the model accuracy. Second, other types of classification algorithms can be considered. Although the machine learning algorithms considered in this study are quite comprehensive, there are still unexplored techniques that can be applied to the prediction problem in the movie domain. . Third, other features or data that may boost the prediction accuracy can be added. For example, movie buzz data on social media such as *Twitter* can be used. We expect that these implementations can be the other ways to improve the prediction of a movie performance.

References

Abe, S. 2005. *Support vector machines for pattern classification*, Springer, London, 58-59.

Adamopoulos, P. 2013. "What makes a great MOOC? An interdisciplinary analysis of student retention in online courses," In *Proceedings of the Thirty Fourth International Conference on Information Systems (ICIS)*, volume 2013, 2013.

Auria, L., and Moro, R. A. 2008. *Support vector machines (SVM) as a technique for solvency analysis*, DIW Berlin, Berlin, Germany.

Asur, S.,and Huberman, B. 2010. "Predicting the future with social media," *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, (1), 492-499.

Balakrishnama, S., and Ganapathiraju, A. 1998. "Linear discriminant analysis - a brief tutorial," *Institute for Signal and information Processing*.

Basuroy, S., Chatterjee, S., & Ravid, S. A. 2003." How critical are critical reviews? The box office effects of film critics, star power, and budgets," *Journal of marketing*, 67(4), 103-117.

Basuroy, S., Desai, K. K., & Talukdar, D. 2006. "An empirical investigation of signaling in the motion picture industry," *Journal of Marketing Research*, 43(2), 287-295.

Blumenthal, H., & Xu, Y. 2012." The ghost club storyscape: designing for transmedia storytelling," *Consumer Electronics, IEEE Transactions on*, 58(2), 190-196.

Chambers, M., & Dinsmore, T. W. 2014. "Advanced Analytics Methodologies: Driving Business Value with Analytics," *Pearson Education*.

DeMaris, A. 1995. "A tutorial in logistic regression," *Journal of Marriage and the Family*, 956-968

De Vany, A., & Walls, W. D. 1999. "Uncertainty in the movie industry: Does star power reduce the terror of the box office?," *Journal of cultural economics*, 23(4), 285-318.

Dhar, T., Sun, G., & Weinberg, C. B. 2012. "The long-term box office performance of sequel movies," *Marketing Letters*, 23(1), 13-29.

Dietterich, T. G. 1997. "Machine-Learning Research," *AI magazine*, 18(4), 97-136.

Du, J., Xu, H., and Huang, X. 2014. "Box office prediction based on microblog," *Expert Systems with Applications*, 41(4), 1680-1689.

Edwards, L. H. 2012. "Transmedia storytelling, corporate synergy, and audience expression," *Global Media Journal*, 12(20), 1-12.

Elberse, A. 2007. "The power of stars: Do star actors drive the success of movies?" *Journal of Marketing* ,71(4), 102-120.

Elder IV, J. F. 2003. "The Generalization Paradox of Ensembles," *Journal of Computational and Graphical Statistics*, 12(4), 853-864.

Eliashberg, J., Elberse, A., and Leenders, M. A. 2006. "The Motion Picture Industry: Critical Issues in Practice, Current Research, and New Research Directions," *Marketing Science*, 25(6), 638-661.

Eliashberg, J., Hui, S. K., and Zhang, Z. J. 2007. From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts," *Management Science*, 53(6), 881-893.

Freund, Y., Schapire, R., & Abe, N. 1999. "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, 14(5), 771-780.

Ghiassi, M., Lio, D., and Moon, B. 2015. "Pre-Production Forecasting of Movie Revenues with a Dynamic Artificial Neural Network," *Expert Systems with Applications*, 42(6), 3176-3193.

Guyon, I., & Elisseeff, A. 2003. "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, 3, 1157-1182.

Han, M.H., Kang, H.M., and Kim, D.S. 2010. "Three Stage Performances and Herding of Domestic and Foreign Films in the Korean Market," *Asia Marketing Journal*, 11(4), 21-48.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. 2009. "The Elements of Statistical Learning: Data mining, Inference and Prediction," *Springer*.

Jenkins, H. 2007. "Transmedia Storytelling 101," http://henryjenkins.org/2007/03/transmedia_storytelling_101.html. Accessed 17 April 2016.

Jenkins, H. 2013. "Transmedia Storytelling: Moving Characters from Books to Films to Video Games Can Make Them Stronger and More Compelling," *MIT Technology Review*, from http://www.technologyreview.com/news/401760/transmedia-storytelling/

Karniouchina, E. V. 2011. “Impact of star and movie buzz on motion picture distribution and box office revenue,” *International Journal of Research in Marketing*, 28(1), 62-74.

Kleinbaum, D. G., & Klein, M. 2010. *Logistic Regression*, New York, Springer, 1-39.

Litman, B. R. 1983. “Predicting success of theatrical movies: An empirical study,” *The Journal of Popular Culture*, 16(4), 159-175.

Liu, Y. 2006. “Word of mouth for movies: Its dynamics and impact on box office revenue.” *Journal of marketing*, 70(3), 74-89.

Long, G. A. 2007. *Transmedia storytelling: Business, aesthetics and production at the Jim Henson Company* (Unpublished doctoral dissertation), Massachusetts Institute of Technology.

Mayr, A., Binder, H., Gefeller, O., and Schmid, M. 2014. “The Evolution of Boosting Algorithms,” *Methods of Information in Medicine*, 53(6), 419-427.

Mitchell, T. M. 1997. *Machine learning*. McGraw-Hill..

Mishne, G., and Glance, N. S. 2006. “Predicting Movie Sales from Blogger Sentiment,” in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 155-158.

Moloney, K. T. 2011. *Porting transmedia storytelling to journalism* (Unpublished doctoral dissertation), University of Denver.

Montillo, A. A. 2009. *Statistical Foundations of Data Analysis*, New York, Springer.

Nelson, R. A., and Glotfelty, R. 2012. “Movie Stars and Box Office Revenues: An Empirical Analysis,” *Journal of Cultural Economics*, 36(2), 141-166.

Perryman, N. 2008. “Doctor Who and the Convergence of Media A Case Study inTransmedia Storytelling,” *Convergence: The International Journal of Research into New Media Technologies*, 14(1), 21-39.

Pohar, M., Blas, M., & Turk, S. 2004. “Comparison of logistic regression and linear discriminant analysis,” *Metodoloki zvezki*, 1(1), 143-161.

Prettenhofer, P., & Louppe, G. (2014). Gradient Boosted Regression Trees in Scikit-Learn. In *PyData, London 2014*

Ravid, S. A. 1999. “Information, Blockbusters, and Stars: A Study of the Film Industry,” *The Journal of Business*, 72(4), 463-492.

Scolari, C. A. 2009. “Transmedia Storytelling: Implicit Consumers, Narrative Worlds, and Branding in Contemporary Media Production,” *International Journal of Communication* (3), 586-606.

Sharda, R., and Delen, D. 2006. “Predicting Box-Office Success of Motion Pictures with Neural Networks,” *Expert Systems with Applications*, 30(2), 243-254.

Simonoff, J. S., and Sparrow, I. R. 2000. “Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers,” *Chance*, 13(3), 15-24.

Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. 2012. “Auto-WEKA: Automated Selection and Hyper-Parameter Optimization of Classification Algorithms.”

Treme, J., and Craig, L. A. 2013. “Celebrity Star Power: Do Age and Gender Effects Influence Box Office Performance?,” *Applied Economics Letters*, 20(5), 440-445.

Yamagishi, J., Kawai, H., & Kobayashi, T. 2008. “Phone duration modeling using gradient tree boosting,” *Speech Communication*, 50(5), 405-415.

Zhang, X. 2003. “Discriminant analysis as a machine learning method for revision of user stereotypes of information retrieval systems,” presented at the *MLIRUM'03: 2nd Workshop Machine Learning, Information Retrieval, and User Modeling, 9th Int. Conf. User Modeling, Pittsburgh, PA*.

Zhang, L., Luo, J., and Yang, S. 2009. “Forecasting Box Office Revenue of Movies with BP Neural Network,” *Expert Systems with Applications*, 36(3), 6580-6587.

Zhu, J., Zou, H., Rosset, S., & Hastie, T. 2009. “Multi-class adaboost,” *Statistics and its Interface*, *2*(3), 349-360.