How to Predict Financial Success of Movies based on Big Data Analytics

Yifan Yao
Mechanical Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania, 15213, U.S.A.

Keywords: Analytic, Financial Movie, Machine Learning, Big Data, Prediction.

Abstract: The movie industry is a field with high potential values but movies' success is highly unpredictable. In this study, a model is developed to predict the financial performance of movies using big data analytics methods, which can serve as a reference for movie decision makers. The box office is considered as the major indicators of the financial performance of movies. Six features are defined to describe the movies and seven machine learning-based algorithms are used to predict the box office of movies. Four models are selected among them as component models to formulate the cinema ensemble model (CEM) by voting the estimates of four component models. The accuracy of the CEM model is tested by the APHR test, which is about 58.5% of Bingo rate. As a result, the accuracy of the CEM model is higher than all of the seven single component models.

1 INTRODUCTION

Nowadays, the movie industry serves as a profitable field of media all around the world, which produces billions of dollars of revenue every year. However, the cost of making a movie is also considerable. Movie makers do not always make a profit but often result in a loss. Research shows that the average investment to produce a movie is 65 million USD. Between 2000 and 2010, only one-third of movies in the US were profitable (Murschetz et al. 2020). Numerous cases of failed movies illustrate the high risk of movie-making industry, where people are always uncertain whether a movie will be highly profitable or completely failed. As Lehrer et al. mentioned, improved forecasts are valuable because they could increase capital investments by reducing investor uncertainty of the box office consequences and also help marketing teams tailor effective advertising campaigns (Lehrer, Xie 2021).

Therefore, it is critical for movie investors to develop an approach to predict the success of a movie before the movie is produced, in order to make better decisions about whether or not it is worthwhile to invest money in it. Additionally, for cinemas, predicting the success of movies is also of significant need because releasing an unsuccessful movie means a loss in box-office revenue, which is the main profit source of cinemas. Therefore, in this study, a model is introduced to predict the success of movies before investors and cinemas make financial decisions about the movies. The cinema ensemble model consists of 4 component models out of 7 machine learning-based models, and the result is voted by all the component models. The introduced CEM model has relatively high accuracy in predicting the box office level of a movie before it is released. The model can provide useful information for movie donators and cinemas to make proper decisions.

2 DATA ACQUISITION

The analyzed dataset is collected from the Korean Film Council webpage and naver.com. This dataset includes top the 400 movies by their number of views, which are released from October 25, 2012, to December 31, 2014. The movies that go beyond the top 400 are not considered in this study, because according to the classification standard presented in Table 1, most of those movies are classified into class 6 (Flop Class). Therefore, classifying those movies will be way much easier, which will increase the accuracy of prediction of our results, but the improvement is not appreciated since it will only decrease the effectiveness of our evaluation. In

Yao, Y.
How to Predict Financial Success of Movies based on Big Data Analytics.
DOI: 10.5220/0011304800003437
In *Proceedings of the 1st International Conference on Public Management and Big Data Analysis (PMBDA 2021)*, pages 238-243
ISBN: 978-989-758-589-0

addition, film production companies and cinemas are usually not interested in predicting the performance of movies that have a limited budget, but are only interested in the movies that have more budget which will hopefully create a large amount of revenue. Hence, including movies beyond the top 400 is not worthwhile. Among those collected movies, the author discarded 25 movies that miss the important information for the features evaluated in this study, and left the rest of 375 movies. The distribution of classes for those data are presented in Figure 1 (Lee et al. 2018).

Table 1: Movie performance classes.

Class	Attendance Range (in thousands)	Revenue Range (Approx. in $ thousands)
1(Blockbuster)	> 4000	> 26,700
2	2000—4000	13,300—26,700
3	750—2000	5300—13,300
4	250—750	1800—5300
5	100—250	700—1800
6(Flop)	< 100	< 700

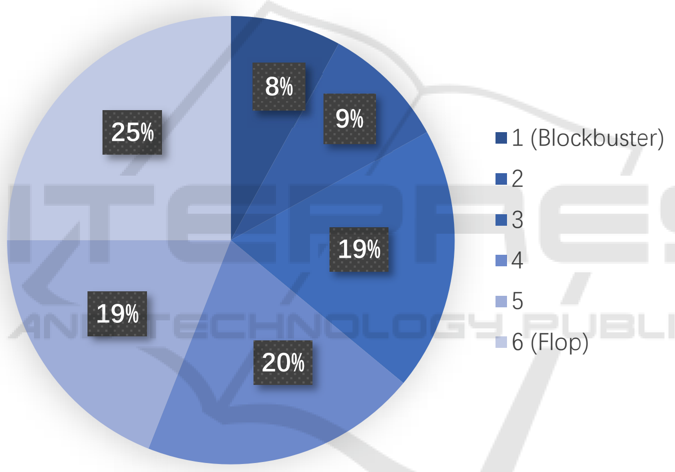


Figure 1: Distribution of movie classes.

3 METHODOLOGY

3.1 Definition

To construct the predictive model of movies, some parameters are needed to be defined to measure the financial success of them. In this study, the box office is considered as the major indicator of movies' financial success, because it directly affects the total revenue a movie produces. In other words, it decides how much money movie investors and cinemas earn if they buy that movie.

The objective of this study is to forecast the box office during the first week after the movie is released, due to the fact that movies obtain the greatest part of box office in the first week. As illustrated in Figure 2, the largest portion of a movie's revenue (40% on average) is obtained from the box office sale during the first week of a release, when other cinema managers are yet to decide whether to show the movie or not (Murschetz et al. 2020).

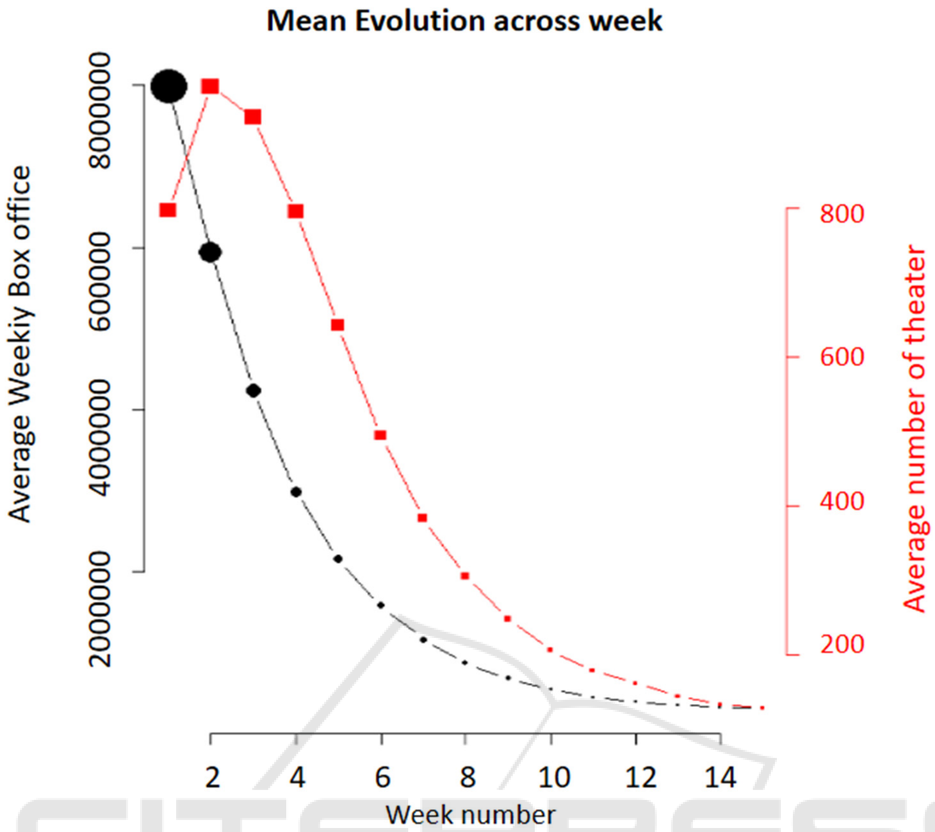


Figure 2: Average weekly box office performance and average number of movie theaters showing a movie.

3.2 Feature Description

3.2.1 Genre

In this study, movies are classified into 16 categories of genres, which include: action, adventure, animation, comedy, criminal, documentary, drama, epic, family, fantasy, horror, independent, mystery, romance, science fiction and thriller. Each movie may subject to multiple categories.

3.2.2 Sequels

Sequels also have some impact on a movie’s success. Producing sequels for released movies is less risky than producing movies with new themes. Dhar et al. 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) (Dhar et al. 2012). Firstly, sequel movies can attract a number of people who have watched previous movies in that series. Secondly, movie investors and cinemas can look at the ratings and box office of previous movies to make sure whether movies in these themes appeal to the customers.

3.2.3 Number of Plays on the Initial Day of Release

The third feature of movies is the number of plays on the initial day of release. The reason why the number of screens is not used as the feature is that the number of plays is affected by the length of movies. For example, a movie with a length of 3 hours will be played fewer times compared to a movie of 1.5 hours provided the same number of screens. Therefore, the shorter movie will have a better chance to succeed in box office. The data of plays at the initial days is collected from the website of KMRB.

3.2.4 Movie Buzz: Before the Release

Movie buzz is the fourth feature that may have an impact on the box office of movies. For example, 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 (Liu 2006). The author counted the number of comments of movies on Naver Movie (see http://movie.naver.com/) in the review section before the movies are released. The number of comments describes how many people were discussing about the movie before it is released.

3.2.5 Transmedia Storytelling

The fifth feature for movies is whether the movie is a transmedia storytelling movie or not. The value 1 means yes and 0 means not. A transmedia storytelling movie is a movie based on television series, comics, or novels, which have been exposed to the public before the movie is released. Remade movies are not considered as transmedia storytelling movies.

3.2.6 Star Buzz

The influence of star power is a strong factor to decide the box office, because superstars always have a number of fans, who will pay for the box office even if the movie is not brilliant enough. Therefore, an indicator is needed to measure the power of a star. The method is to count the number of posts on Naver Blog from two months before the release of the movie and one month before it. That is because if the influence of a star is greater, there will be more people mentioning them on the internet. Also, before the release of a movie, the stars will do some advertisements, which will affect the number of mentions in blogs. Therefore, to avoid this factor, only the posts more than one month are considered before the release when the advertisement has usually not begun yet.

3.3 Algorithms and Models

In this study, 7 different algorithms are used: adaptive tree boosting, gradient tree boosting, linear discriminant, logistic regression, neural networks, random forests, and support vector classifier, to build some candidate classifiers for the dataset. Then all the candidates are compared and the candidates with the high accuracy in prediction are selected as component models. In this paper, a plurality voting system is used in which the winning estimate is the one with the largest votes. Through such a process, an ensemble model for the prediction of a movie's success can be constructed. This model is called Cinema Ensemble Model (CEM). The process is schematized in Fig. 3.

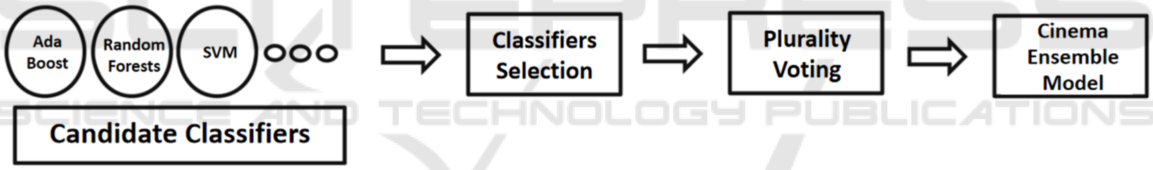


Figure 3: The process to build the CEM model.

4 RESULTS

4.1 Performance Metrics

In this study, the performance metrics of Sharda and Delen (2006) are adopted. Average Percent Hit Rate (APHR) is used to measure the accuracy of their prediction models. There are 2 types of APHRs calculated in this study: Bingo and 1-Away. Bingo counts the number of classifications that exactly matches their actual classes, while 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, it is considered to be a misprediction (Sharda 2006).

4.2 Performance Results

The training process and predictions were done for each of the seven algorithms, and their accuracies were recorded using APHR values. The result is shown in Table 2. Base on the percentage of Bingo and 1-Away values, the performances of models are ranked respectively, shown in Table 3.

Table 2: APHRs of candidate models.

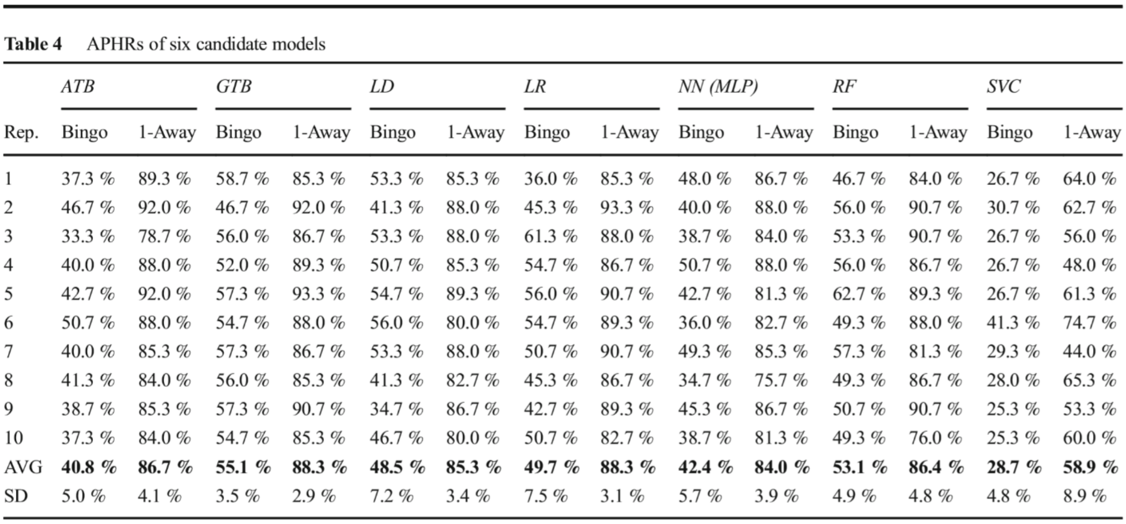
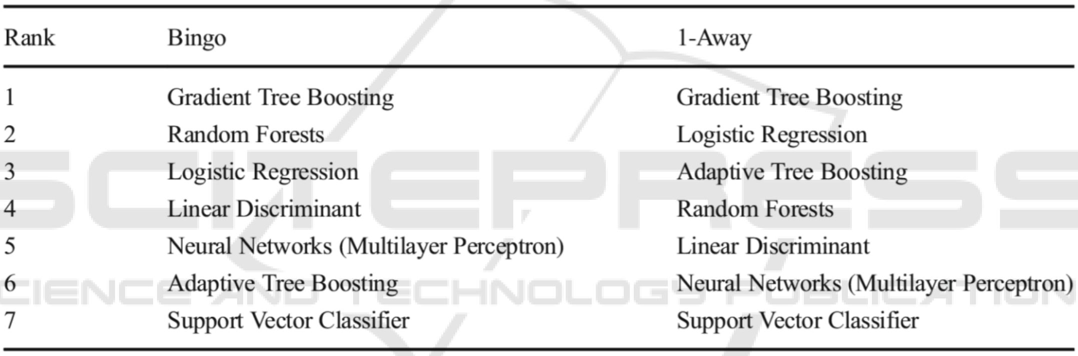


Table 4 APHRs of six candidate models

	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 %

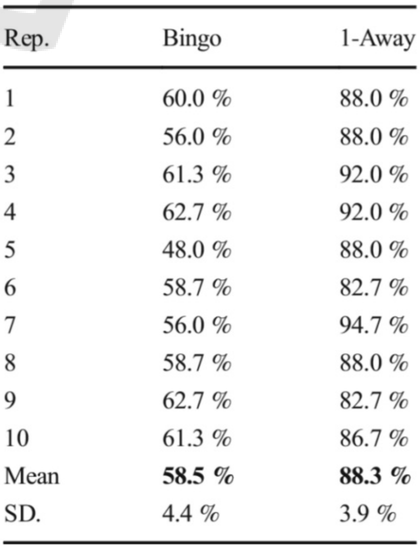
Table 3: Ranking of model performance.



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

It is noticed that gradient tree boosting has 55.1% of Bingo rate and 88.3% of 1-Away rate, both are the highest among seven models. Random forest, logistic regression and linear discriminant also have relatively high Bingo rate and 1-Away rate. Neural networks, adaptive tree boosting and support vector classifier have relatively low Bingo rate. Therefore, GTB, LD, LR and RF are chosen as the four models that have the best performance in predicting the movies' box office. These four models serve as the component models in our CEM model. In the ensemble approach of CEM, the four models vote and the largest vote wins the estimate. Since there are four votes presented, when the number of votes is equal (e.g., two votes A and two votes B), the class that GTB votes will be chosen because GTB has the highest accuracy. The result of the CEM model is shown in Table 4. It has 58.5% of Bingo rate and 88.3% of 1-Away rate, which performs better than any of the seven single models.

Table 4: APHRs of 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 %

5 CONCLUSION

In this study, Cinema Ensemble Model (CEM) is introduced to predict the financial success of a movie. The model is based on machine learning methods and six features of movies: genre, sequel, number of plays on the first day of release, movie buzz before the release, transmedia storytelling and star buzz. Then seven different machine learning classification algorithms are used to predict the level of box office for movies: adaptive tree boosting, gradient tree boosting, linear discriminant, logistic regression, neural networks, random forests and support vector classifier. After evaluating the performance of each model, it is shown that GTB, LD, LR and FR have the best performances, so they are selected to be the component models in the ensemble model (CEM). The four component models estimate the class and the class of the largest number of votes wins the estimation. The result shows that CEM model has 58.5% of accuracy, which is generally higher than previous researches.

This approach of movie box office prediction can be applied by movie investors and cinemas to make financial decisions before movies are released. That is, to decide whether to invest the movie or how much screen will be provided for that movie. It will help them to make more accurate and advisable decisions and therefore produce more revenue for the film industry.

For researches in the future, firstly, more kinds of component models can be used and more times of trials can be applied to increase the accuracy of the CEM model. Secondly, more features can be used to identify the movies. For example, sentiment analysis for dialogues in movies can be applied to describe the structure of movies and the peaks in movies.

REFERENCES

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

Lee, K., Park, J., Kim, I. et al. Predicting movie success with machine learning techniques: ways to improve accuracy. Inf Syst Front 20, 577–588 (2018). https://doi.org/10.1007/s10796-016-9689-z.

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

Murschetz, P. C., Bruneel, C. ., Guy, J.-L. ., Haughton, D. ., Lemercier, N. ., McLaughlin, M.-D. ., Mentzer, K. ., Vialle, Q. ., Zhang, C. ., Murschetz, P. C. ., & Bakhtawar, B. (2020). Movie Industry Economics: How Data Analytics Can Help Predict Movies' Financial Success. Nordic Journal of Media Management, 1(3), 339–359. https://doi.org/10.5278/njmm.2597-0445.5871.

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications,30(2), 243-254.

Steven F. Lehrer, Tian Xie (2021) The Bigger Picture: Combining Econometrics with Analytics Improves Forecasts of Movie.