Movie box office prediction based on ensemble learning

Shuangyan Wu
School of Computer
South China Normal University
Guangdong 510970, China
20175031@m.scnu.edu.cn

YuFan Zheng
Nanfang College of
Sun Yat-Sen University
Guangdong 510970, China

Zhikang Lai
Nanfang College of
Sun Yat-Sen University
Guangdong 510970, China

Fujian Wu
Nanfang College of
Sun Yat-Sen University
Guangdong 510970, China

Choujun Zhan
School of Computer
South China Normal University China
zchoujun2@gmail.com
Corresponding Author

***Abstract*—The movie box office is now considered a relatively unpredictable short-term experience product. The profits of the film industry are constantly expanding, and more and more investors are engaged in it. But its uncertainty has caused huge losses for many investors. In this paper, film data from 1980 to 2018 were collected on box office mojo, and then, we use machine learning methods, including the Ensemble learning algorithm, to build a predictive model. Results show that the gradient boosting decision tree (GBDT) gives the best performance, of which R_2 is higher than 0.995. Experimental results show that the Ensemble learning algorithm is much better than the traditional machine learning algorithm.**

***Index Terms*—Ensemble learning , Movie box office prediction ,Correlation coefficient**

I. Introduction

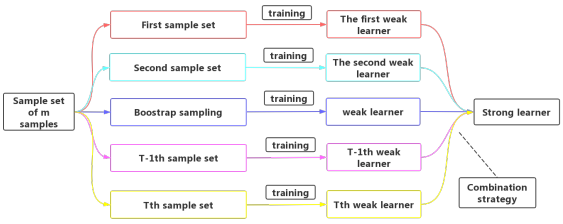
Movie has a history of more than 100 years and become an indispensable part of world culture. Now, it is not only an important object for people to entertain and relax, but also an important medium for cultural exchanges between different countries and regions. In the modern world, movie has also become a business with huge market profit and potential. Hence, it attracts many investors each year and everywhere in the world. The box office is the most important source of income for the film industry. The prediction of the box office becomes a critical issue. Accurate box office prediction can help adjust the sales strategy of the movie after release according to the predicted box office, so as to maximize the profit brought by the box office. However, big investment may not have big output, and the lack of box office prediction tools makes investors unable to effectively avoid investment risks. Therefore, the prediction of the box office is of great significance. Compared with traditional model, machine learning model has better prediction effect.

Until now, there have been relatively few studies on the prediction of movie box office. In 1981, Austin B A analyzed the American Film Institute's film grading system to see if the film's rating would have a potential impact on the film audience [1]. In 1989, Litman B R et al. utilize regression analysis to study the factors related to the success of the film industry and found that actors, characters, stories and industry associations was a key factor [2]. In 1994, the Sochay S's box office prediction model is developed, which used 22 features to predict film rentals and movie running time. The final evaluation result R^2 is 0.380 [3]. In 2003, King T [4] studied the relationship between the 2003 American film score and the total box office, and found that the correlation between the film's score and the box office income is zero. However, for movies with more than 1,000 screens released, the score was positively correlated with the total box office. In 2006, Sharda R et al. classified films based on movie box office revenues and used neural networks to predict the category of movies [5]. Duan W et al. analyzed the box office through the word of mouth of films on the Internet and found that the most significant influence of box office correlation was the number of word of mouth of films [6]. In 2009, Zhang L et al. divided all films into six grades based on box office revenue, in order to predict the correct level of the film. They used a multi-layer BP neural network to build the model [7]. Additonally, Zhang W et al. proposed the box office prediction based on news reports. They used regression analysis and KNN model based on IMDB's movie data and Lydia's news data. Finally, it is found that news data plus movie data can get a better prediction result [8]. In 2013, Oh C used the box office data from Twitter and box office mojo to analyze, and found that not only the number of word of mouth can affect the box office, but also the participation of consumers and the content generated by marketers could indirectly influence box office [9]. Mestyán M et al. predicted the box office of the movie based on Wikipedia activities [10]. They also used a multiple linear regression model to predict the box office of the movie. More specifically, they use data from collective activities to make predictions. The popularity of a movie is predicted by measuring and analyzing the activity levels of editors and viewers of corresponding entries in the movie.

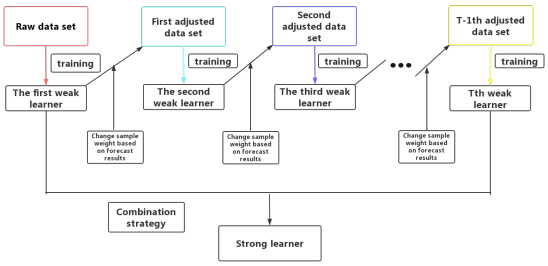
In this article, we used the movie data set collected from

the Box office mojo for developing prediction model for the Box office. Here, we used 21 features to predict the box office. Ensemble learning method, such as GBDT and Random Forest (RF) model, and traditional machine learning algorithm decision trees (DT) are used to establish the prediction model. Evaluation criterion, including MSE, RMSE, MAE, MAD, R^2, Adjusted R^2, were used to evaluate the model. This paper is arranged as following: Section II briefly introduces the ensemble learning algorithm and several evaluation indicators used in the box office prediction model. In Section III, the data and explains the experimental process.The fourth part is the comparison between the Ensemble learning algorithm and the traditional machine learning algorithm DT have been introduced. Finally, in Section IV, the experiment of this paper and the future proposes improvements are discussed.

II. Machnine Learing Models



(a) The flowchart of bagging method.



(b) The flowchart of boosting method.

Fig. 1: Illustration of bagging and boosting.

Bagging and Boosting are two common ideas to improve machine learning methods.

- Boosting, which uses a number of weak learners to create a strong learner, is an idea widely used in Ensemble learning. The principle of the Boosting algorithm is to give priority to the training set to give the same weight, and train a weak learner according to the training set (Fig. 1(a));
- Bagging (Bootstrap Aggregating) uses Boostrap sampling for the training data, that is, the data is sampled back. Each time the sampled data is utilized to train a weak learner. T weak learners are obtained after T re-sampling sampling, and then a strong learner is synthesized according to the combined strategy weak learner (Fig. 1(b)).

A. Gradient Boosting Decision Tree

Gradient Boosting Decision Tree (GBDT) is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. GBDT uses a forward distribution algorithm in which weak learners use only CART. The purpose of the alternate iteration of GBDT is to find a weak learner such that the loss function is smaller than the loss function obtained by the weak learner of the previous iteration. Therefore, it is necessary to use CART of each iteration to fit the negative gradient of the loss function on the weak learner obtained in the previous iteration. This ensures that the overall loss of the model is declining.

B. Random Forest

Random forests (RF) are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. In the regression problem, the weak learner always is CART. Usually, when the subtree of CART is divided, it will find the best feature among all the features to divide. The difference of RF is that it firstly randomly selects some features, and then finds the optimal feature from these features to divide. This enhances the generalization capability of the model.

C. Evaluation Criterion

In this study, y represents the real box office of a movie, while f stands for the forecast box office. Additionally, let $\overline{y}$ represents the average of the box office of a movie. y_i and f_i represent the i^{th} week/day true value and the predicted value, respectively. Suppose there are n samples, each with k features. Here, we adopt the following criterion to evaluate the results.

- Mean Absolute Error (MAE) : MAE is the average of the absolute errors between all predicted and true values, and can visually reflect the error between them.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|. \tag{1}$$

- Mean Square Error (MSE) : The MSE is the average of the sum of the squares of the errors between all predicted and true values. It is a commonly used indicator for evaluating regression tasks.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2. \tag{2}$$

- Root mean square deviation (RMSE) : RMSE is essentially the root number for MSE. It is to better understand the error of the predicted value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2}. \tag{3}$$

- Coefficient of determination(R^2) : R^2 represents the degree of fitting of the predicted value to the curve formed by the true value.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n} (y_i - f_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}. \quad (4)$$

where $SS_{res} = \sum_{i=1}^{n} (y_i - f_i)^2$ and $SS_{tot} = \sum_{i=1}^{n} (y_i - \overline{y})^2$. Its range is between (-1, 1). The closer to 1 the better the degree of fit.

- Adjusted R^2 : Adjusted R^2 is an extended criterion based on R^2 to solve the inaccuracy of the evaluation value caused by the increase in the value of R^2 as the number of samples increases. If a useful variable is added to the model, Adjusted R^2 will increase and vice versa.

$$R^2_{adj} = 1 - \left[\frac{\left(1 - R^2\right)(n-1)}{n-k-1} \right]. \quad (5)$$

- Median absolute error (MAD) : MAD is a robust regression assessment indicator. It is more adaptable to outliers in the results than RMSE and MSE.

$$MAD = median_i \left(|y_i - f_i| \right), \quad (6)$$

where $median_i(.)$ represents the median value among the i-th values.

D. Correlation coefficient

The Pearson coefficient product moment correlation coefficient which is used to measure the degree of correlation (linear correlation) between two variables X and Y,

$$\begin{aligned} COV(X,Y) &= \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right), \\ COR(X,Y) &= \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}}. \end{aligned} \quad (7)$$

Note that Pearson correlation coefficient can only explain the degree of linear correlation between variables.

III. Experiment results

This work focus on building a predictive model for predicting the box office of a movie. A dataset including almost all the movies released by American distributors, such as Warner Bros, Universal, Buena Vista, Fox, etc from 1980 to 2018 is adopted in this study. This dataset contains detailed information on 13,373 movies, which includes "movie title", "daily gross" or "weekly gross", "rank", "Budget", "Theater", "gross oversea" etc. Based on this data set, we adopt 21 features for developing model to predict movie box office.

Since the "movie runtime" and "movie budget" are partially missing, we removed these two features in the regression prediction. Then, we investigate the Person correlation between the variable to be predicted y and all the other features x_i by calculating the correlation coefficients between them (shown in Table I). Ensemble learning models, including RF and GBDT, are utilized to predict the box office. We compare the performance of these methods with the traditional machine learning prediction model, such as DT. In order to obtain a reliable prediction result, we conducted multiple random cutting training sets and test sets with a ratio of 8:2 for multiple experiments. We use the grid search method to find the optimal hyperparameters by adopting different hyperparameters combination. In order to obtain a reliable prediction result, we also conducted multiple random cutting training sets and test sets with a ratio of 8:2 for multiple experiments. Then, we obtained the average evaluation value of each prediction models.

- First, we adopt seven features with correlation coefficients greater than 0.9 to establish prediction models. Table II shows the experimental results;
- Then, we adopt all the 21 features to establish predict models. Table III shows the experimental results.

TABLE I: Correlation coefficient between the values to be predicted y and the selected feature x_i.

Feature	Correlation coefficient
Movie Runtime	0.282827
Movie Budget	0.682980
Day1 DailyGross	0.619987
Day2 DailyGross	0.640083
Day3 DailyGross	0.684663
Day4 DailyGross	0.717395
Day5 DailyGross	0.713146
Day6 DailyGross	0.696471
Day7 DailyGross	0.666679
Day1 GrossToDate	0.946997
Day2 GrossToDate	0.962766
Day3 GrossToDate	0.972609
Day4 GrossToDate	0.979207
Day5 GrossToDate	0.984519
Day6 GrossToDate	0.988866
Day7 GrossToDate	0.992406
Day1 DailyTheater	0.662555
Day2 DailyTheater	0.666456
Day3 DailyTheater	0.669568
Day4 DailyTheater	0.672648
Day5 DailyTheater	0.675915
Day6 DailyTheater	0.679452
Day7 DailyTheater	0.682958

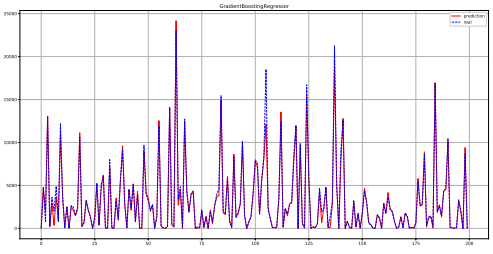


Fig. 2: prediction results of a movie.

Fig. 2 shows the prediction result of a movie. The $y-$axis is the box office value, while x-axis the time. The read line is the predicted value and the true box office. Obviously, the prediction results have a good fit with the real results, which

indicates that the GBDT model predictions have achieved good results.

TABLE II: Experiment results based on model with seven features.

	GBDT	RF	DT
MAE	154.42996	155.507787	209.611033
MSE	194013.463	188610.379	338780.887
MAD	51.8720946	54.1481238	64.3959539
R^2	0.994654391	0.994800846	0.990527654
R^2_{adj}	0.994651524	0.994792473	0.990522575
RMSE	425.400073	421.1147916	572.990988

TABLE III: Experiment results based on model with 21 features.

	GBDT	RF	DT
MAE	149.496681	143.116421	210.535717
MSE	172573.201	189605.432	401742.741
MAD	44.5121321	44.2699923	62.2137375
R^2	0.995191244	0.994783589	0.988905743
R^2_{adj}	0.995183500	0.994775188	0.988887876
RMSE	409.770810	423.409948	619.835795

Table II and III show that the ensemble learning model GBDT and RF performs better than the traditional machine learning model DT. Obviously, experimental results show that R^2 is greater than 0.9, which indicates that the model fits the historical data well. Comparing the predictive results provided by models based on 21 features and models based on 7 features, we find that R^2, MAE and MAD of models based on 21 features are better. However, the other criterion, MSE and RMSE are worse. This phenomenon shows that although the complexity of the model is increased, the robustness of the model is improved.

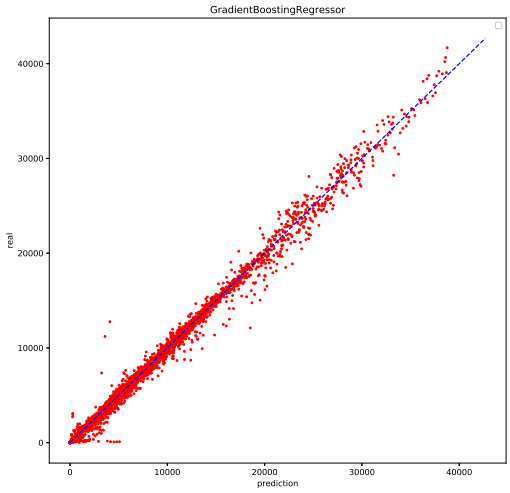


Fig. 3: Prediction results provided by the GBDT.

The prediction results provided by the GBDT is shown in Fig. 3. y-axis represents the true value, while x-axis is the predicted value. The blue line stands for the prediction value equals to the true value. We found that the model predictions are all near the blue line. Larger deviations occurred for samples with larger and smaller box office values. However, the overall experimental results are satisfactory.

IV. Conclusion

Prediction is one of the most important issues for studying human's behavior [11], [12]. This paper introduces the use of the ensemble learning model, RF and GBDT, to predict the movie box office, while using the Pearson correlation coefficient for selecting model features. We find the best R^2 is greater than 0.995 and had a better performance than the decision tree. In the future work, we plan to use the popular deep learning models LSTM and GRU to build prediction models. Additionally, we want to built a generic model that will classify movies in accordance with MAPP ratings, and then build predictive models for each MPAA-rated movie. After that, we will classify the movie according to the box office data and build a model to predict the success of the box office.

Acknowledgment

This work was supported by Science and Technology Program of Guangzhou (201904010224) and National Science Foundation of China (61703355).

References

[1] B. A. Austin, "The influence of the mpaa's film-rating system on motion picture attendance: A pilot study," *The Journal of Psychology*, vol. 106, no. 1, pp. 91–99, 1980.

[2] B. R. Litman and L. S. Kohl, "Predicting financial success of motion pictures: The'80s experience," *Journal of Media Economics*, vol. 2, no. 2, pp. 35–50, 1989.

[3] S. Sochay, "Predicting the performance of motion pictures," *Journal of Media Economics*, vol. 7, no. 4, pp. 1–20, 1994.

[4] T. King, "Does film criticism affect box office earnings? evidence from movies released in the us in 2003," *Journal of Cultural Economics*, vol. 31, no. 3, pp. 171–186, 2007.

[5] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," *Expert Systems with Applications*, vol. 30, no. 2, pp. 243–254, 2006.

[6] W. Duan, B. Gu, and A. B. Whinston, "The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry," *Journal of retailing*, vol. 84, no. 2, pp. 233–242, 2008.

[7] L. Zhang, J. Luo, and S. Yang, "Forecasting box office revenue of movies with bp neural network," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6580–6587, 2009.

[8] W. Zhang and S. Skiena, "Improving movie gross prediction through news analysis," in *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 301–304, IEEE, 2009.

[9] C. Oh, "Customer engagement, word-of-mouth and box office: the case of movie tweets," *International Journal of Information Systems and Change Management*, vol. 6, no. 4, pp. 338–352, 2013.

[10] M. Mestyán, T. Yasseri, and J. Kertész, "Early prediction of movie box office success based on wikipedia activity big data," *PloS one*, vol. 8, no. 8, p. e71226, 2013.

[11] Q. Wen, C. Zhan, Y. Gao, X. Hu, E. Ngai, and B. Hu, "Modelling human activity with seasonality bursty dynamics," *IEEE Transactions on Industrial Informatics*, 2019.

[12] B. Mei, X. Wang, Q. Wen, Y. Tang, H. Wang, and C. Zhan, "A novel algorithm for estimating purchase incentive of the public based on mobile cloud computing," *IEEE Access*, 2019.