ISSN 2616-7433 Vol. 6, Issue 7: 12-16, DOI: 10.25236/FSST.2024.060703

Reliability Analysis of Multiple Machine Learning Methods in Influencing Factors Analysis and Prediction of War Film Box Office——Coordination and Optimization between Narrative Rhythm and Box Office Performance of War Film

Yonghao Hong

School of Liberal Arts, Nanjing University, Nanjing, Jiangsu, 210000, China

Abstract: *The influence factors of the war film box office are divided into two parts according to external factors and internal components, with a total of 15 types of variables. The regression effect of war film box office is compared based on 10 machine learning regression models, it is found that Random Forest model shows high reliability. According to the concept of econometrics, the narrative rhythm of war film can be measured qualitatively through the quantitative expression of 8 internal elements. It is found that there is a correlation between narrative rhythm and film box office. Taking 8 elements as independent variables and narrative rhythm coordination degree as dependent variable, 10 kinds of machine learning classification are carried out, and it is found that AdaBoost model is the most reliable.*

Keywords: *machine learning; war film; influencing factors; narrative rhythm; commercialization*

1. Introduction

Academia has been put generous effort on the prediction of film box office based on machine learning models, but the data indicators of training set and test set are inferior(Wang Cui and Zhang Haiyue 2019[1]; HeQi and Yuan Fangying 2021[2];Ye Na *et al*.,2022[3]). From the contents, it can been found that these studies only consider about external factors to research the factors affecting box office. However the internal components are also important factors. Except the external factors, This paper is committed to investigating application of machine learning model in the internal components of war film box office. The emergence of cinemetrics provides a possibility for quantitative research on the narrative rhythm of films. According to Yang Shizhen's viewpoint, the cinemetrics focuses on introducing econometric methods and big data technology into film research, bringing a new theoretical paradigm and research methods to traditional film research(Yang Shizhen 2019). [4]

2. Materials and Methods

2.1 Data Sources and Specifications

Based on the definition of war film in the 2005 "*Dictionary of Film Art*", this study defines war film as a non science fiction feature film based on historical reality, which shows the war military operations taken by an independent nation or country to resist or invade, strive for national independence or annex other countries. Based on the online film review and the author's film viewing, it is found that war films with box office less than 10million yuan generally lack data. Therefore, the war films released in Chinese mainland from January 1st, 2000 to May 1st, 2024 and with a box office of more than 10million yuan in the Chinese film market were selected for investigation. Based on this, 48 war films were collected.

This study sets 15 independent variables. According to the nature, divide them into external factors and internal components. The former includes: *nature*, *number of participating stars*, *well-known director*, *holiday film*, *history portrayed*, *pre-release publicity*, *post-release publicity*.The latter includes: *film time*, *the length of the first battle from the beginning of the film*, *the average length of the battle*

ISSN 2616-7433 Vol. 6, Issue 7: 12-16, DOI: 10.25236/FSST.2024.060703

clips, *the median length of the battle clips*, *the number of battle clips*, *the proportion of battle clips*, *the average interval length of battle clips*, *the median interval length of battle clips*. One dependent variable is *film box office*.The 15 independent variables are identified as follows: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O.

Among these independent variables, the *nature* is divided into *action* and *campaign*. According to the war time and space shown in these films, the *history portrayed* can be divided into 9 types: *Ancient times*, *War to Resist US Aggression and Aid Korea*, *War of Resistance Against Japan*, *European Battlefield*, *Ten-Year Civil War*, *the Pacific War, War of liberation*, S*uppressing Bandits*, *Anti Japanese Struggle*.

The data for external factors come from the Mao Yan Pro app and internal components come from the author's recording. Among the 15 variables, variables related to time are all accurate to 1 second. In data processing, for film box office data, unit is expressed in yuan. Raw data is used in the statistics, while natural logarithms are used in operations to represent it. For categorical data, convert it into dummy variables for operation.

2.2 Research Methods and Steps

Machine learning is the use of computer technology to perform computational analysis on collected data, and continuously improve computational methods to enhance the accuracy of completing specific systems(Meng Ziliu and Li Tenglong,2020). [5] However, machine learning regression and classification can't obtain definite equations like traditional models, and the model is usually evaluated by testing the classification performance of the data.

This study intends to use the machine learning regression method in the SPSS PRO statistical analysis modeling platform, Decision Tree, Random Forest, Adaboost, CatBoost, GBDT, KNN, Neural Networks, SVR, XGBoost, Extra Trees, totaling 10 pairs, the data are divided into training set and testing set in a 7:3 ratio.

In the optimization algorithm, Heuristic Algorithm (Genetic Algorithm) is used to find the optimal hyperparameters for each model. In data training, the model is trained by shuffling. Finally, establish table based on indicators, such as R^2, MSE, RMSE, MAE, MAPE to evaluate the regression performance of each model.

3. Data Aggregation and Visualization Analysis

Based on the R^2 value in Table 1, it can be judged that among the 10 machine learning regression models, only Random Forest has strong reliability, while other machine learning models are poor, especially Neural Network and SVR have obvious shortcomings in handling the box office prediction problem of war film.

Table 1 Model Evaluation

Data Set	Index	DT	RF	AdaBoost	CatBoost	GBDT	KNN	ANN	SVR	XGBoost	ExtraTree
training set	R^2	1	0.926	0.999	1	1	0.065	0.568	-3.293	1	0.926
	MSE	0	0.135	0.002	0	0	2.19	0.887	9.746	0	0.173
	RMSE	0	0.368	0.044	0.01	0	1.48	0.942	3.122	0.013	0.416
	MAE	0	0.294	0.011	0.008	0	1.191	0.681	2.569	0.006	0.35
	MAPE	0	1.58	0.064	0.046	0	6.475	3.684	14.374	0.027	1.881
testing set	R^2	0.606	0.897	0.572	0.727	0.562	0.099	-0.218	-2.445	0.485	0.836
	MSE	1.146	1.039	0.695	0.852	0.749	1.495	3.657	9.083	0.553	0.588
	RMSE	1.07	1.019	0.834	0.923	0.512	1.248	1.912	3.014	0.744	0.667
	MAE	0.904	0.888	0.728	0.628	2.677	6.719	1.598	2.59	0.571	0.599
	MAPE	4.872	2.772	3.985	3.924	0.598	2.234	8.561	15.697	3.165	3.154

According to the importance ratio of each variable displayed by Random Forest, it can be seen that among the external influencing factors of the film, the *post-release publicity* is 31.9%, the *number of participating stars* is 28.2%, and the *holiday film or not* is relatively low. The importance of *holiday film* is less than 1%. See Table 2 .

ISSN 2616-7433 Vol. 6, Issue 7: 12-16, DOI: 10.25236/FSST.2024.060703

Table 2 Feature Importance Ratio(%)

External Factors							Internal Components							
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1.0	28.2	5.6	0.3	3.2	1.7	31.9	6.3	2.6	2.3	5.0	3.6	1.7	4.8	1.8

4. Analysis of the Relationship between the Narrative Rhythm and Commercialization of War Film

4.1 The Composition of Narrative Rhythm in War Film

As of now, from the perspective of academic research on war narratives, more research has focused on narrative types, narrative strategies, and narrative modes, while research on narrative rhythm is still lacking. In fact, not only the narrative rhythm of war films, but also the narrative rhythm of all films can't be quantitatively reflected. More importantly, it is based on the viewing subject's perception of the film, resulting in a compact or loose overall rhythm. However, this qualitative expression still can't clarify the specific differences reflected by the different rhythms between different films.

Therefore, it is believed that from the perspective of metrology the internal constituent elements in war film can be used to quantitatively represent the narrative rhythm of war film: the primary task of war films is to showcase war. The narration often depicts the harsh reality of war... showing the disasters and emotional trauma that war brings to humanity on the screen(Yang Teng and Xu Ming,2019).[6]

However, the narrative rhythm of war film, as a variable that cannot be quantified by data or objectively evaluated, is composed of non time series and non normal distribution dependent variables. Therefore, linear or nonlinear regression cannot be used to construct a specific function for expressing the fast or slow narrative rhythm of war film. In this case, the author proposes a new measurement concept: the narrative rhythm of war film is not qualitatively measured by speed, but by coordination or non coordination. Therefore, using the coupling coordination model in SPSSAU, evaluate whether the narrative rhythm of 48 war films is imbalanced:

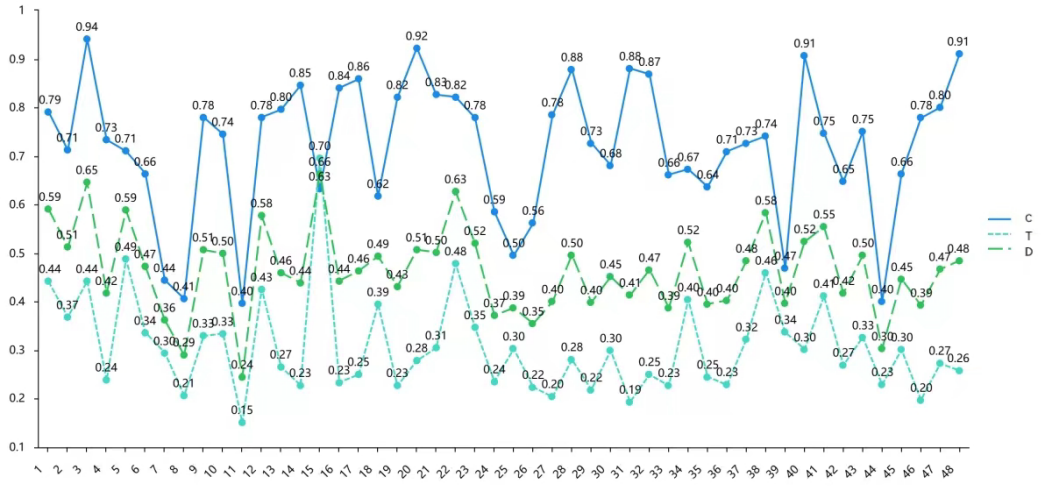

Figure 1 Coupling Coordination Degree

From Figure 1, it can be seen that there are 14 coordination items and 34 dissonance items, with coordination accounting for 29.17%. Based on the requirements of the "14th Five Year Plan" issued by the Chinese government for the development goals of films, the box office revenue of films is divided into 100 million yuan. Among the 24 war films with box office revenue greater than or equal to 100 million yuan, there are 10 coordination items and 14 imbalance items, accounting for 41.67% of the total. From the data, it can be subjectively inferred that although there are still imbalances in high box office films, overall they perform better in terms of narrative rhythm coordination.

Is there a correlation between the coordination of narrative rhythm in war film and the high or low box office of war film from an objective perspective? And can multiple machine learning classification

ISSN 2616-7433 Vol. 6, Issue 7: 12-16, DOI: 10.25236/FSST.2024.060703

algorithms be used to achieve coordinated optimization of the narrative rhythm of war film based on the internal constituent elements of war film? Further verification is required.

4.2 Verification of the Relationship between the Narrative Rhythm of War Film and Box Office

Considering that the correlation between the narrative rhythm coordination of war film and the box office of war film cannot be judged using correlation analysis methods such as Spearman Correlation Analysis and Pearson Correlation Analysis, the Grey Relational Analysis in SPSSAU is used for research.

Grey Relational Analysis is a multi factor statistical analysis method that assists decision-making by studying the degree of correlation between the reference sequence and the feature sequence. The specific steps are as follows: set *Ln_ film box office* as the reference sequence.Set the *coupling coordination degree D* value of war film narrative rhythm as the feature sequence, then label the variable with the letter Q. Use the Mean Method to perform dimensionless processing on the two variables, with a distinguishing coefficient of 0.50. The results are shown in Table 3:

Table 3 Correlation Results

Item	Correlation Degree	Generalized-Related-Degree Coefficient		
Q	0.664	absolute correlative degree	incidence correlative degree	comprehensive correlative degree
		0.5011	0.5061	0.5036

From Table 3, it can be inferred that there is a correlation between the coordination of narrative rhythm and box office in war film. Based on Figure 1, the importance of each variable that constitutes the coordination of narrative rhythm is greater than 0. Accumulation reveals that the importance of coordination of narrative rhythm accounts for 28.1%. But if the 8 variables that make up the elements are integrated into the 1 type variable of *war narrative rhythm* coordination, and combined with the other 7 types of variables into machine learning regression, find that the R^2 values of the 10 machine learning regressions do not exceed 0.6, indicating poor performance, which should be caused by a small number of features. In summary, the coordination of narrative rhythm can be considered an important indicator that affects box office performance to a certain extent. But when making box office regression prediction, it needs to be split into 8 variables.

4.3 Coordinated Classification of Narrative Rhythm in War Film Based on Multiple Machine Learning Models

It has been shown that there is a correlation between the coordination of narrative rhythm in war films and the high box office of war films, machine learning classification can be used to optimize and adjust the narrative rhythm of war film. Therefore, whether the narrative rhythm of war films is coordinated is set as the dependent variable, 8 types of internal constituent elements are set as independent variables for 10 machine learning classification models. The data source and specifications, research methods and steps are consistent with the first part. Finally, the accuracy, recall, precision, F-score are organized and establish a table to evaluate the classification performance of each model.

Table 4 Model Evaluation Form

Data Set	Index	DT	RF	AdaBoost	CatBoost	GBDT	KNN	ANN	SVR	XGBoost	ExtraTree
training set	R^2	1	0.926	0.999	1	1	0.065	0.568	-3.293	1	0.926
	MSE	0	0.135	0.002	0	0	2.19	0.887	9.746	0	0.173
	RMSE	0	0.368	0.044	0.01	0	1.48	0.942	3.122	0.013	0.416
	MAE	0	0.294	0.011	0.008	0	1.191	0.681	2.569	0.006	0.35
	MAPE	0	1.58	0.064	0.046	0	6.475	3.684	14.374	0.027	1.881
testing set	R^2	0.606	0.897	0.572	0.727	0.562	0.099	-0.218	-2.445	0.485	0.836
	MSE	1.146	1.039	0.695	0.852	0.749	1.495	3.657	9.083	0.553	0.588
	RMSE	1.07	1.019	0.834	0.923	0.512	1.248	1.912	3.014	0.744	0.667
	MAE	0.904	0.888	0.728	0.628	2.677	6.719	1.598	2.59	0.571	0.599
	MAPE	4.872	2.772	3.985	3.924	0.598	2.234	8.561	15.697	3.165	3.154

From Table 4, it can be seen that except for neural network, the other 9 machine learning classification models can classify the coordination of narrative rhythm in war films. The F-score of AdaBoost model classification is all 1, showing the best performance.

ISSN 2616-7433 Vol. 6, Issue 7: 12-16, DOI: 10.25236/FSST.2024.060703

5. Conclusion

Based on the above data results, the following conclusions can be drawn:

Firstly, based on 10 machine learning classification models, a war film box office regression model is established. Combining the performance of each model in the training and testing sets, it is found that the Random Forest model has strong reliability in predicting war film box office.

Secondly, the combination of cinemetrics theory and machine learning method has solved the shortcomings of traditional film commercialization path analysis in handling complex data relationships and dealing with multiple influencing factors. It has improved the predictive effect of previous studies that only considered external factors as variables, providing a more reliable method for predicting box office of war film.

Thirdly, the narrative rhythm of war film can be qualitatively measured by coordination, which is based on quantitative data and has objectivity and accessibility compared to fast or slow qualitative standards.

Fourth, film producers should pay attention to multiple promotions after the film is released and use of celebrities to participate in the film after their release, which can significantly improve box office performance. In addition, under the qualitative criteria of whether the narrative rhythm is coordinated, attention should also be paid to the coordination of the narrative rhythm in war film. Before making a film, AdaBoost classification model can be used for optimization to increase box office revenue.

Acknowledgememt

This paper is the follow-up research result of the author's award-winning project (No.20243124) at the 27th Forum of Sciences & Arts of Nanjing University.

References

[1] Wang Cui, Zhang Haiyue,2019.Research on the Relevance between Film Content Elements and Box Office Based on Machine Learning and Natural Language Processing Algorithms.Advanced Motion Picture Technology (09),4-9.

[2] He Qi,Yuan Fangying,2021.Study on the Influencing Factors of Film Consumption and Box Office Prediction in the Context of Digital Transformation.Price:Theory & Practice (09),163-167+204. DOI:10.19851/j.cnki.CN11-1010/F.2021.09.296.

[3] Ye Na, Xu Xin, Li Chengtong, Chai Zhenhui, Fu Xinghong,2022.Research on influencing factors and prediction of film box office based on machine learning.Public Communication of Science & Technology (22),89-92+96.DOI:10.16607/j.cnki.1674-6708.2022.22.022.

[4] Yang Shizhen,2019.Theory, Method and Application of Cinemetrics.Contemporary Cinema (11),32. (in chinese)

[5] Meng Ziliu, Li Tenglong,2020.Review and Prospect of Machine Learning Technology. Application of IC.37(10),56-57.doi:10.19339/j.issn.1674-2583.2020.10.024.(in chinese)

[6] Yang Teng, Xu Ming,2019.On the narrative mode of war movies.Movie Literature (05),104-106.(in chinese)