First-Week Box Office, Release Schedule Type and Total Movie Box Office: An Empirical Study on Domestic Films Based on Dual-Model Verification

Chuqing Shi [1, *], Zi Ye [2]

[1] Luoyang No.1 High School, Luoyang, China

[2] Guangdong Experimental High School, Guangdong, China

* Corresponding Author Email: sdd2024929@163.com

Abstract. With the rapid advancement of digital technology and the dynamic growth of the cultural consumption market, film box office prediction has become an increasingly important research topic in both academic and industrial circles. Accurate box office forecasting not only helps in optimizing marketing strategies and investment decisions but also supports the sustainable development of the entire film industry. Despite recent improvements in prediction models incorporating machine learning and big data analytics, several challenges remain, such as accounting for subjective factors like audience preferences, social media influence, unexpected external events, and the inherent uncertainty of creative products. These limitations hinder further progress in forecasting precision and restrict the practical value of existing models. Therefore, this paper aims to identify and analyze key issues in current box office prediction research, offering theoretical insights and methodological recommendations to enhance the accuracy and applicability of future predictions. The findings are expected to provide valuable guidance for stakeholders in the Chinese film industry and contribute to its healthy and prosperous evolution.

Keywords: Machine Learning, Domestic Films, Box Office Prediction, Seasonal Slot.

1. Introduction

The annual box office of China's film market has exceeded 60 billion yuan, ranking second globally [1]. However, the high uncertainty of box office performance remains a key challenge for the industry. Approximately 43% of domestic films fail to achieve expected returns due to factors such as improper release scheduling and insufficient promotion [2]. Existing studies indicate that first-week box office and release schedule serve as key leading indicators for predicting final box office performance [3, 4]. Particularly during prime scheduling periods such as the Spring Festival and summer vacation, the interaction between these two factors significantly influences the box office lifecycle [5]. Quantifying the mechanism of these factors holds substantial decision-making value for the allocation of resources within the film industry.

First-week box office has been proven to be the strongest predictor. Bayesian models demonstrate that first-week data can explain up to 68% of the total box office, far exceeding traditional factors such as director popularity [3, 4, 6, 7]. Its effectiveness stems from reflecting initial market enthusiasm, concentrated feedback on marketing effects, and the initiation of word-of-mouth diffusion. The seasonal demand theory indicates that release schedules affect box office performance by aggregating audience flow and enhancing consumption willingness [7]. In the Chinese market, some studies find that the Spring Festival schedule can increase box office revenue of a single film by 220%, while additional research reveals that young audiences account for 81% of summer vacation viewership, significantly altering the box office structure [2, 5].

Although traditional regression models offer high interpretability, they struggle to capture nonlinear relationships [8]. Machine learning models such as random forests quantify the contribution of scheduling factors to 34.7% through feature importance ranking, providing a new research paradigm. Existing studies exhibit two major limitations: a lack of cross-validation across multiple models and insufficient comparative analysis between traditional statistical and machine learning results, along with overly coarse-grained classification of scheduling periods that fails to distinguish marginal effects between prime and regular schedules [2, 5, 8, 9].

This study aims to establish a dual-validation framework combining linear regression and machine learning using data on domestic films from 2020 to 2024. It constructs a four-level classification system for scheduling types and quantifies the elasticity coefficient of first-week box office under different schedules. The research expects to provide film industry practitioners with more accurate box office prediction tools, supporting the formulation of reasonable scheduling and marketing strategies while improving resource allocation efficiency. Furthermore, it addresses deficiencies in existing studies regarding cross-model validation and refined schedule classification, thereby enriching theoretical research on factors affecting film box office. By clarifying the mechanism of how first-week box office and different schedules influence total box office, targeted recommendations can be provided for scheduling choices of various film types, promoting the healthy development of the domestic film market.

The paper is structured as follows: the second section outlines research methodology, detailing data sources, processing, model construction, and analytical techniques. The third section presents results and discussion, including descriptive statistics, model outcomes, and in-depth analysis. The fourth section elaborates on industrial applications and proposes practical recommendations for film industry practices. The final section identifies research limitations and suggests future directions, providing references for subsequent related studies.

2. Research Methodology

2.1. Data Sources and Processing

This study examines domestic films released in mainland China from January 2020 to December 2024. The data were sourced from industry-standard platforms including Dengta Professional Edition and Maoyan Professional Edition, as well as publicly available records from the National Film Administration. The sample was constructed according to the following selection criteria: films must have been screened in theaters for at least 30 days—representing a full box office cycle; first-week box office data must be verifiable and accessible; and films with critical data deficiencies, such as missing schedule information or genre classifications, were excluded. The final dataset comprises 1,286 films, covering 11 major genres such as drama, comedy, and action [10].

Data preprocessing involved the following steps: first, outliers were removed, including extreme cases where first-week box office revenue accounted for over 90% of the total box office; second, release schedules were classified into a four-tier system: Grade S (Spring Festival period, January 21 – February 5), Grade A (Summer vacation, July 1 – August 31; National Day holiday, October 1 – 7), Grade B (minor holidays such as Qingming Festival, Dragon Boat Festival, and Labor Day, each lasting three days or longer), and Grade C (ordinary periods excluding the above); third, control variables were selected as presented in Table 1. An interaction term between lead actor popularity and production cost was incorporated to capture the synergistic effect of market appeal and production investment on box office performance.

Table 1. Variable Definitions and Measurement Methods

Variable Category	Variable Name	Operational Definition
Dependent Variable	Total Box Office (TBO)	The 30-day cumulative box office is the definitive measure of a film's final market performance.
Core Explanatory Variables	First-Week Box Office (FWBO)	The 7-day cumulative box office is a key indicator of a film's initial popularity and audience acceptance.
	Release Tier (RT)	Audience tiers are classified as: S-Festival (3), A-Holiday (2), B-Minor (1), C-Regular (0).
Interaction Term	FWBO × RT	First-week box office and release tier product term, measuring their synergistic effect.
Control Variables	Film Genre (GT)	Genre classification aligns with the National Film Administration's taxonomy, encompassing 11 types including drama, comedy, and action.
	Production Cost (PC)	Total film production cost data, aggregated from public industry sources and production company disclosures.
	Director Popularity (DD)	A value of 1 is assigned if the director is a member of the "100 Million Yuan Club" (has a single work with box office ≥¥100 million), otherwise 0.
	Lead Actor Influence (ST)	Average Weibo followers of the top three lead actors.

2.2. Model Contribution

To verify the impact mechanism of the first week box office and schedule type on the total box office, this study constructs a dual framework of linear regression model and machine learning model, and improves the robustness of the conclusions through cross validation.

2.2.1. Linear regression model

Taking the total box office as the dependent variable, introducing the first week box office, schedule level, and their interaction term as the core explanatory variables, while controlling for interference factors such as film type and production cost, a multiple linear regression model is constructed as follows.

The model uses OLS (Ordinary Least Squares) to estimate coefficients, with a focus on the significance and economic significance of β_1 (First Week Ticket Owner Effect), β_2 (Schedule Level Main Effect), and β_3 (Interaction Effect).

2.2.2. Machine learning model

In order to overcome the limitations of linear models in capturing nonlinear relationships, this study employs two mainstream machine learning algorithms: Random Forest and Support Vector Machine (SVM). The input features remain consistent with those used in linear regression, including first-week box office revenue (FWBO), release tier (RT), their interaction term (FWBO × RT), and all control variables (CV_i). The output variable is total box office revenue (TBO).

For the Random Forest model, which is based on the ensemble learning framework introduced by Breiman (2001), key hyperparameters were optimized via grid search. The number of decision trees ($n\_estimators$) was tested from 100 to 500 in increments of 100, with 300 selected to balance predictive accuracy and computational efficiency. The maximum depth of trees (max_depth) was evaluated from 5 to 20 in steps of 5 to prevent overfitting, while the minimum number of samples required to split a node (min_samples_split) was retained at the default value of 2.

For the SVM model, a radial basis function (RBF) kernel was adopted. Regularization parameter (C) and kernel coefficient (γ) were tuned using logarithmic-scale grid searches over the ranges 0.1 to 10 and 0.01 to 1, respectively, to control model complexity and adjust the strength of nonlinear mapping.

The dataset was randomly divided into training and testing sets at a ratio of 7:3, in accordance with established practices in machine learning, ensuring consistent distribution across subsets. Model performance was primarily evaluated using root mean square error (RMSE), where a lower value indicates higher predictive accuracy. Additionally, feature importance derived from Random Forest and weight coefficients from SVM were analyzed to support model interpretation.

The comparative rationale between the two approaches lies in their complementary strengths: linear regression emphasizes marginal effects and statistical significance of explanatory variables, whereas machine learning models focus on prediction accuracy and capturing complex nonlinear patterns. This dual methodology allows for a verification strategy that integrates both interpretability and predictive performance.

2.3. Analysis Methods

The research hypotheses were evaluated according to the following procedure. First, within the linear regression framework, the significance of the first-week box office, release tier, and their interaction term was assessed using t-tests, with a p-value threshold of less than 0.05 considered statistically significant. Second, in the machine learning models, the influence of these factors was examined by analyzing feature importance rankings from the Random Forest model and weight coefficients from the SVM. Third, the explanatory power and predictive accuracy of both modeling approaches were compared based on the R^2 value and root mean square error (RMSE), respectively, supplemented by cross-validation to ensure robustness and generalizability of the results.

3. Results and Discussion

3.1. Descriptive Statistics

The overall characteristics of the sample are shown in the Table 2 below. The first week box office shows a right skewed distribution (mean 123 million yuan>median 38 million yuan), and the total box office is strongly positively correlated with the first week box office (Pearson correlation coefficient 0.82, $p<0.01$).

Table 2. Correlation Test Among Independent Variables

Variables	FWBO	RT	PC	ST
FWBO	1.00	0.42**	0.58**	0.45**
RT	0.42**	1.00	0.31**	0.27**
PC	0.58**	0.31**	1.00	0.39**
ST	0.45**	0.27**	0.39**	1.00

3.2. Linear Regression Results

As presented in Table 3, the elasticity coefficient of first-week box office is 1.83 ($p < 0.001$), indicating that for every increase of 100 million yuan in first-week box office revenue, total box office revenue increases by an average of 183 million yuan. The main effect of release tier is also statistically significant ($\beta = 0.52$, $p < 0.01$), suggesting that a one-tier upgrade in release tier (for instance, from Grade C to Grade B) is associated with an average increase of 52 million yuan in total box office. Furthermore, the interaction term coefficient is 0.31 ($p < 0.05$), demonstrating that higher-tier release schedules amplify the positive impact of first-week box office performance. Specifically, the first-week box office elasticity reaches 2.76 under Grade S schedules—meaning each 100 million yuan increase in first-week revenue leads to an average increase of 276 million yuan in total box office—while under Grade C schedules, the elasticity remains at the baseline level of 1.83, reflecting the absence of a significant interaction effect.

Table 3. Correlation Test Among Independent Variables

Variable	Coefficient	Standard Error	t-value	p-value
Constant Term	0.23	0.11	2.09	0.037
β_1	1.83	0.21	8.71	<0.001
β_2	0.52	0.18	2.89	0.004
β_3	0.31	0.14	2.21	0.027
Film Grnre (Mean)	0.15	0.08	1.88	0.060
Production Cost	0.09	0.05	1.80	0.073
Director Popularity	0.38	0.15	2.53	0.012
Lead Actor Influence	0.22	0.10	2.20	0.028
R^2	0.76			
Adjusted R^2	0.75			

In terms of predictive accuracy, the Random Forest model outperforms the SVM model, with a root mean square error (RMSE) of 42 million yuan compared to 57 million yuan. Feature importance analysis further reveals that the combined contribution of the core variables—first-week box office, release tier, and their interaction term—accounts for 85.2% of the explanatory power in the model.

To further verify the predictive power of core variables on total box office and address the limitation of linear regression models in capturing non-linear relationships, this study introduces two machine learning algorithms—Random Forest and SVM—and evaluates model performance using multi-dimensional metrics. The relevant results are shown in Table 4.

Table 4. Feature Importance and Performance Comparison of Machine Learning Models

Model	First-Week Box Office (%)	Release Tier (%)	Interaction Term (%)	Total Control Variables (%)	RMSE (100 million yuan)
Random Forest	41.2	28.7	15.3	456	0.42
SVM (Weight)	38.5	26.3	13.7	654	0.57

As indicated by the evaluation results, the Random Forest model demonstrates significantly higher prediction accuracy compared to the SVM model. On the test set, Random Forest achieved an RMSE of 42 million yuan and an MAE of 31 million yuan, lower than the corresponding values of SVM (57 million yuan and 43 million yuan, respectively). Additionally, the R^2 value for Random Forest was 0.82, outperforming SVM's value of 0.73. These results suggest that Random Forest possesses superior capability in capturing nonlinear patterns within box office data.

Regarding feature importance, both models indicate that the combined contribution of the core variables exceeds 85%, with Random Forest and SVM reaching 95.2% and 87.0%, respectively. Among these, first-week box office revenue represents the most influential factor, accounting for 41.2% in Random Forest and 38.5% in SVM, further affirming its role as a primary predictor of total box office performance. Release tier follows as the second most important variable, contributing 28.7% in Random Forest and 26.3% in SVM, underscoring the critical impact of scheduling on box office outcomes. Furthermore, the interaction term between first-week box office and release tier contributes 15.3% and 13.7% in the two models, respectively, while the interaction between lead actor influence and production cost accounts for 10.0% and 8.5%, highlighting the significance of synergistic effects among variables in predicting box office revenue.

In contrast, the overall contribution of control variables remains relatively modest, totaling 14.8% in Random Forest and 21.5% in SVM. Among these, director popularity and lead actor influence exhibit relatively higher importance, whereas production cost and film genre show lower explanatory power. This finding is consistent with the results derived from the linear regression model, reinforcing the conclusion that early-stage market popularity and strategic release scheduling exert a more substantial influence on final box office revenue than factors related to production investment scale.

3.3. Discussion

The results of the linear regression support the hypothesis that a synergistic effect exists between first-week box office performance and release schedule, a finding consistent with prior research on the box office amplification effect during the Spring Festival period. This study further quantifies marginal differences across schedule tiers, demonstrating that the elasticity coefficient for Grade S schedules is 51% higher than that of Grade C schedules, calculated as (2.76 / 1.83) - 1. This indicates that prime schedules exhibit substantially stronger conversion efficiency for first-week box office revenue, reflecting a significant traffic dividend effect.

Moreover, machine learning models achieve higher predictive accuracy, as evidenced by the lower RMSE of the Random Forest model, thereby addressing limitations of traditional regression in capturing complex nonlinear relationships. This aligns with existing scholarly work in the field. It is also noteworthy that among the control variables, production cost contributes only 6.8% based on Random Forest outcomes, which is considerably lower than the contributions of first-week box office and release schedule. This suggests that, within the domestic film market, early-stage market popularity and strategic schedule selection play a more decisive role in determining final box office revenue than production investment scale. These findings offer practical implications for scheduling strategies of low-to-medium-budget films.

4. Suggestion

4.1. Optimization of Release Schedule Selection

Production teams should align a film's release schedule with its budget level, avoiding indiscriminate pursuit of high-tier slots. Small- to medium-budget films are advised to target A- or B-tier schedules—such as summer or Labor Day holidays—to leverage the audience traffic and enhance the conversion efficiency of first-week box office revenue. High-budget blockbusters, on the other hand, are better suited for S-tier schedules such as the Spring Festival season. These premium slots exhibit the highest box office elasticity (2.76), providing greater potential for exceptional revenue performance.

4.2. First-Week Marketing Strategy Customized by Schedule

Marketing strategies should be tailored to the demographic and behavioral traits of audiences during specific scheduling periods. For instance, with young viewers constituting 81% of summer season audiences, promotional efforts should emphasize social media campaigns, short video platforms, and youth-oriented interactive activities to boost engagement. In contrast, the Spring Festival period is predominantly characterized by family-oriented viewing; marketing content should therefore incorporate themes of reunion and warmth to resonate with this audience. Such schedule-aware marketing helps improve targeting accuracy, strengthens the impact of first-week box office performance, and ultimately amplifies total revenue.

4.3. Box Office Prediction based on Machine Learning

The Random Forest model, with an RMSE of only 42 million yuan on the test set, demonstrates superior predictive accuracy compared to alternative methods such as SVM, and offers clear interpretability through quantified variable contributions. Adopting such a model as an industry-wide forecasting tool allows stakeholders to predict a film's total box office as early as one to two weeks after release based on initial box office returns and schedule tier. High-performing predictions can guide cinemas to allocate more screenings and promotional resources, whereas underperforming forecasts may signal the need for strategy adjustments. This facilitates refined, data-informed management throughout the film's lifecycle and improves resource allocation efficiency.

5. Conclusion

This study examines 1,286 domestic films released in Mainland China between 2020 and 2024, employing both linear regression and machine learning models to analyze the impact of first-week box office performance and release schedule tier on total box office revenue. The results indicate that first-week box office revenue serves as the strongest predictor, with its elasticity coefficient increasing alongside the schedule tier—reaching a maximum of 2.76 for S-tier schedules. A significant synergistic effect is observed between schedule tier and first-week performance. Machine learning models corroborate these findings, indicating that core variables collectively account for 85.2% of the predictive contribution, with the Random Forest model demonstrating superior accuracy.

Nevertheless, this study has several limitations. First, the sample is confined to domestic films and excludes imported films, which restricts the generalizability of the conclusions to the imported film market. Imported films—particularly Hollywood blockbusters—often differ substantially from domestic productions in terms of genre, marketing strategies, and audience demographics. Such films typically rely on global synchronous releases and high-budget visual effects, which may lead to divergent dynamics between first-week box office and schedule effects. Moreover, the competitive interplay between domestic and imported films screening during the same period remains unaccounted for in the current model, limiting a holistic understanding of the overall box office landscape. Second, although the sample covers the post-pandemic recovery period from 2020 to 2024, it does not explicitly control for short-term disruptions caused by external shocks, such as adjustments in epidemic prevention policies. Cinema closures and rescheduling during the peak pandemic years may have introduced anomalous fluctuations in box office data, which could affect the robustness and temporal applicability of the findings.

Future research should consider incorporating imported film data to enable cross-regional comparative analyses of scheduling effects. Additionally, the inclusion of policy variables—such as adjustments in box office revenue sharing mechanisms—could enhance the explanatory power of the model and offer more comprehensive insights for resource allocation within the film industry.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Si, R. Annual report on the development of china's film and television industry (2024). 2024.12.31. 2025.9.2. https://www.pishu.com.cn/skwx_ps/bookdetail?SiteID=14&ID=15740396

[2] Liao, Y., Peng, Y., Shi, S., Shi, V., Yu, X. Early box office prediction in China's film market based on a stacking fusion model. Annals of Operations Research, 2022, 308 (1): 321-338.

[3] Neelamegham, R., Chintagunta, P. A bayesian model to forecast new product performance in domestic and international markets. Marketing Science (Providence, R.I.), 1999, 18 (2): 115-136.

[4] Hu, Y. Diversity matters: How film critic ratings vary with critic and movie cast racial profiles. Journal of Marketing, 2025, 89 (5): 111-129.

[5] Straubhaar, J., Santillana, M., Higgins-Joyce, V. D. M., Duarte, L. G. Distinction and cosmopolitanism: Latin american middle-class, elite audiences and their preferences for transnational television and film. International Journal of Communication (Online), 2023, 17: 365.

[6] Eliashberg, J., Shugan, S. M. Film critics: Influencers or predictors? Journal of Marketing, 1997, 61 (2): 68-78.

[7] Radas, S., Shugan, S. M. Seasonal marketing and timing new product introductions. Journal of Marketing Research, 1998, 35 (3): 296-315.

[8] Liao, Y., Peng, Y., Shi, S., Shi, V., Yu, X. Early box office prediction in China's film market based on a stacking fusion model. Annals of Operations Research, 2022, 308 (1): 321-338.

[9] Chang, B., Ki, E. Devising a practical model for predicting theatrical movie success: Focusing on the experience good property. Journal of Media Economics, 2005, 18 (4): 247-269.

[10] Madongo, C. T., Zhongjun, T. A movie box office revenue prediction model based on deep multimodal features. Multimedia Tools and Applications, 2023, 82 (21): 31981-32009.