A Multimodal Machine Learning Framework for Box Office Prediction Using CNN-Based Trailer Analysis and Random Forest Models

Xinyuan Li[1, a], Zeyu Yang[2, *], Lei Yue[3], Tianyu Zhao[4]

[1] Northwest Agriculture and Forest University, Yangling, Shaanxi, China

[2] Northwest Agriculture and Forest University, Yangling, Shaanxi, China

[3] Northwest A&F University, Yangling, Shaanxi, China

[4] Northwest A&F University, Yangling, Shaanxi, China

[a] 1725430551@qq.com

[*] Corresponding author

Abstract

This study presents a comprehensive multimodal machine learning framework for box office prediction in the Chinese film market. We develop an integrated approach combining multiple computational techniques: time series analysis for market trend visualization, matrix heatmap analysis for actor-genre correlation mapping, convolutional neural network (CNN) for quantitative sentiment analysis of movie trailers through image processing, and random forest ensemble learning for final prediction. Our methodology employs web crawling technology to collect comprehensive datasets from major film platforms. The CNN model achieves 98.21% accuracy in classifying emotional tones from trailer frames, while the random forest model demonstrates robust performance with MAPE of 10.7% on validation data. We validate our framework by predicting box office performance for three films scheduled for 2025 release, achieving relative errors below 20%. This research contributes to the intersection of computer vision, machine learning, and entertainment industry analytics, providing a scalable solution for revenue forecasting in digital media markets.

Keywords

Machine learning, convolutional neural network, random forest, data mining, computer vision, sentiment analysis, ensemble learning, multimodal analysis, prediction models, feature extraction.

1. Introduction

1.1. Research Background

Film serves as an important vehicle for cultural soft power and value communication. The Chinese government's cultural powerhouse strategy emphasizes promoting high-quality cultural industry development, positioning film as a key component in transitioning from cultural product output to civilizational dialogue. The rapid growth of the Chinese film market reflects enhanced spiritual and cultural needs, creating a relationship that reshapes cultural consumption structure.

As cultural consumption capacity increases, films have become composite cultural carriers satisfying emotional and cognitive needs. Film box office prediction plays a crucial role in quality assessment, helping directors, screenwriters, and investors better evaluate content and depth. Domestic film box office prediction is significant for China's cultural confidence path, contributing to sustainable industry development and facilitating Chinese cinema's transition from industrial output to civilizational dialogue.

1.2. The Specific Role of Box Office Prediction in Cultural Communication

Box office prediction's core value lies in guiding film production quality upgrades through scientific forecasting and promoting healthy industrial ecosystems, supporting cultural confidence strategy through domestic traditional cultural communication and international cultural confidence display. In domestic communication, films decode traditional culture through modern narrative grammar, transforming abstract values into visualized emotional experiences and enhancing cultural identity by reinterpreting traditional concepts in contemporary contexts.

For international display, high-quality film content shapes China's national image globally, achieving shifts from symbolic representation to value resonance. The Wandering Earth series exemplifies this through outstanding presentation, breaking cross-cultural barriers and demonstrating contemporary Chinese cinema's innovative vitality. By embodying universally shared human values, such films foster cultural resonance and promote Chinese civilization dissemination, strengthening China's cultural soft power reach.

2. Sources of Data and Analytical Methods

2.1. Sources of Data

Annual box office revenues and film release numbers in China are sourced from the China Box Office website (boxofficecn.com). Training and testing datasets for the film box office prediction model are derived from the Top 300 Chinese films by total box office revenue on Douban (douban.com). The evaluation and analysis data for film trailers are obtained from trailer videos published on Maoyan (maoyan.com). Our team collected and summarized data from 220 domestic Chinese films released in recent decades, establishing a comprehensive dataset for model development and validation.

2.2. Statistical and Analytical Methods and Models

This study employs four main model types: time series fitting and forecasting model for analyzing domestic Chinese film trends; matrix heatmap model for analyzing actor-genre correlations to determine actor suitability for specific genres; CNN (Convolutional Neural Network) model for sentiment analysis of movie trailers, quantifying emotional values; and random forest model for multifactorial analysis to predict future box office performance, validated by predicting box office for three films scheduled for 2025. We assigned values to 22 features of the collected films using CNN model and heatmap matrix analysis, as shown in Fig. 1. Subsequently, we established a random forest model to train and test the obtained data, ultimately making predictions for three films scheduled for 2025 release.

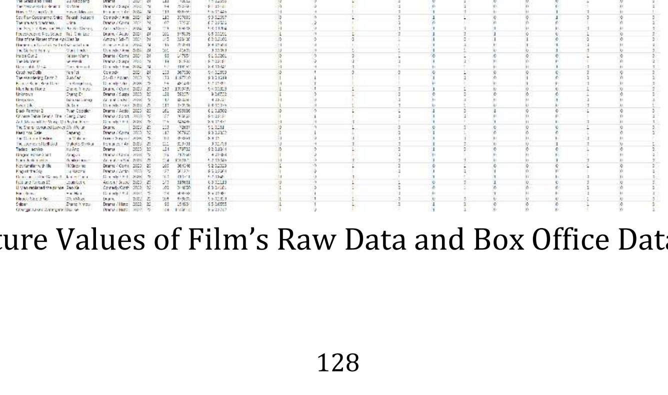

Fig. 1 Feature Values of Film's Raw Data and Box Office Data (Excerpt)

3. Establishment and Analysis of Box Office Prediction Models

3.1. Time Series Fitting and Trend Forecasting

The time series fitting model employs mathematical or statistical methods to model chronologically arranged data, capturing inherent patterns such as trends, periodicity, and noise for forecasting future values or elucidating underlying data dynamics. The objective is to identify optimal parameters enabling the model to best fit historical data and generalize to future predictions [1]. Based on historical trends in domestic film box office revenues and the potential influence of holidays on box office performance, we used MATLAB software to collect data from the China Box Office website.

We gathered data on the number of films released annually from 1996 to 2023, total box office revenue, and calculated average box office revenue for each year. The time series fitting model charts reveal an upward trend in all metrics, with notable growth rate increases after 2012 (see Fig. 2). This trend indicates China's film industry is experiencing rapid development as the nation progresses and quality of life improves. By leveraging time series fitting, we can anticipate the future trajectory of China's film market, inferring that film viewing demand is rising and film is increasingly becoming a form of spiritual and cultural sustenance as the economy grows and society advances.

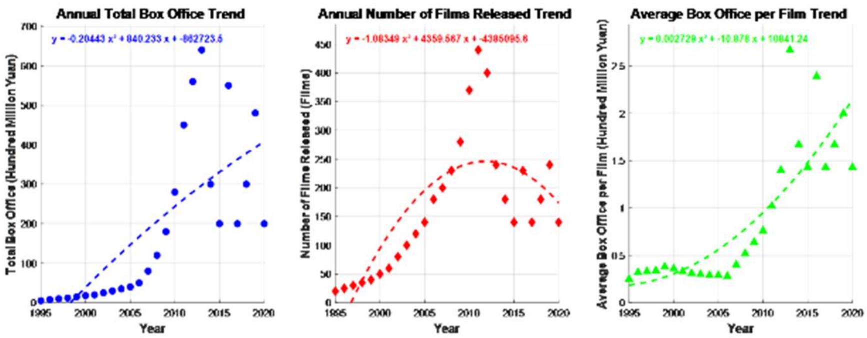

Fig. 2 Analysis of Box Office Market Trends (Total Annual Box Office, Number of Releases, Average Box Office)

3.2. Construction of the Heatmap Matrix

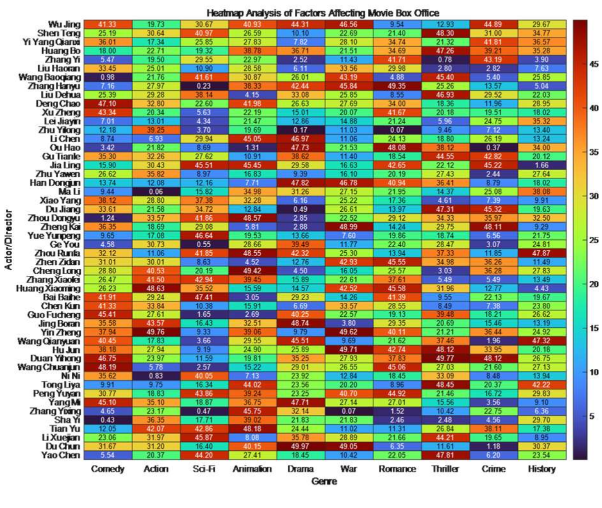

Fig. 3 Heatmap of Correlation between between Box Office Performance of Film Genres and Actors

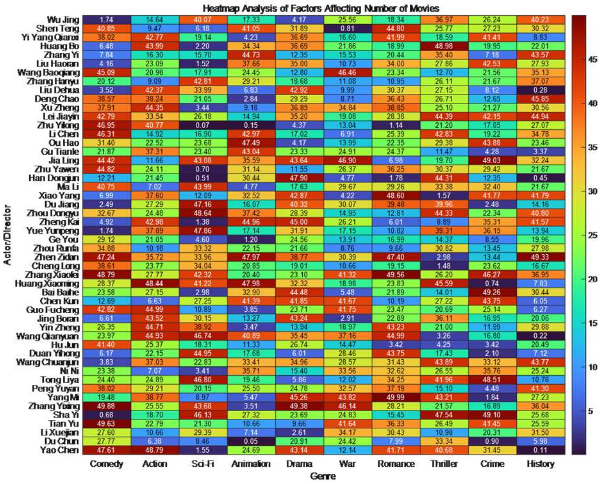

Fig. 4 Heatmap of the Correlation between the Number of Film Genres and Actors

A heatmap matrix is a data visualization technique used to display the relative density or distribution of data within a given area [2]. It typically employs color mapping to represent the density or magnitude of data values. In a heatmap, each data point or region is assigned a color value that reflects the level of data concentration or the numerical value at that specific location. Therefore, we collected the information of the top 50 influential leading actors and the film genres they have acted in from the authoritative film review website Douban ((douban.com)). We then analyzed the representative actors of each film through a heatmap. In the heatmap, the darker the color, the more pronounced the correlation. Through the heatmap, we can clearly see the compatibility between actors and film genres (see Fig. 3). When the number of film genres is greater than or equal to 3, we consider it as 1 (otherwise as 0), and use it as the dependent variable in the feature values (see Fig. 4).

3.3. Introduction, Construction, and Analysis of CNN

A CNN (Convolutional Neural Network) is a multi-layer neural network, where each layer consists of multiple two-dimensional feature maps. As the number of layers increases, the dimensions of the feature maps decrease in size, while the number of feature maps within the same layer increases. This process involves downsampling and feature extraction. Eventually, a one-dimensional vector is obtained, which is fully connected to the output layer [3].

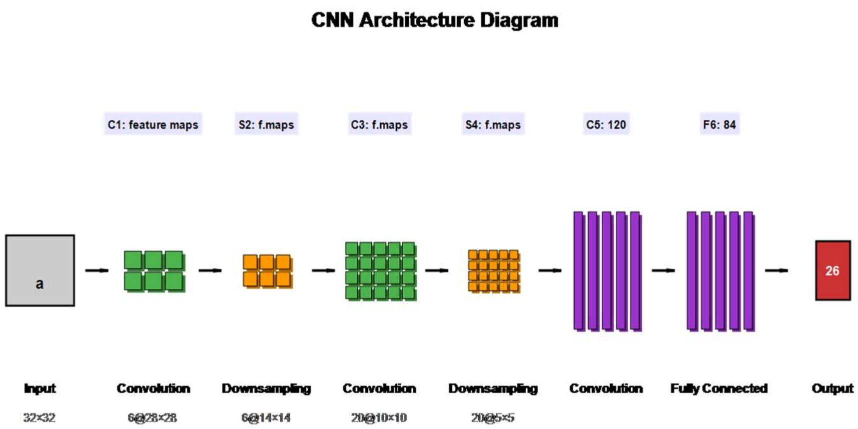

Fig. 5 Convolutional Neural Network Architecture [4]

Fig. 5 illustrates the structure of a CNN. The network comprises several simple and complex neurons, denoted as S-units and C-units, respectively. Different combinations of S-units form S-maps, and different combinations of these S-maps constitute an S-layer. C units are structured in the same manner. Each C-layer is immediately followed by an S-layer, with the two connected in series. The input layer of the network consists of a single image, which serves as the visual input of the network and can directly take the image to be processed as input. C-layers and S-layers are alternately connected, with the feature maps of the final C-layer converging to a single point, forming a one-dimensional vector. This vector is fully connected to the output vector, thus completing the construction of a CNN.

Based on the PyTorch runtime environment, we have written a Convolutional Neural Network model using the Python language. The code example is shown in Fig. 6:

```python
import torch.nn as nn
from torch.nn import functional as F

class SEBlock(nn.Module):
    def __init__(self, channel, reduction=16):
        super().__init__()
        self.avg_pool = nn.AdaptiveAvgPool2d(1)
        self.fc = nn.Sequential(
            nn.Linear(channel, channel // reduction),
            nn.ReLU(inplace=True),
            nn.Linear(channel // reduction, channel),
            nn.Sigmoid()
        )

    def forward(self, x):
        b, c, _, _ = x.size()
        y = self.avg_pool(x).view(b, c)
        y = self.fc(y).view(b, c, 1, 1)
        return x * y.expand_as(x)

class simplecnn(nn.Module):
    def __init__(self, num_class=2):
        super().__init__()
        self.features = nn.Sequential(
            nn.Conv2d(3, 32, kernel_size=3, padding=1),
            nn.BatchNorm2d(32),
            nn.LeakyReLU(0.1),
            nn.MaxPool2d(2),
            nn.Conv2d(32, 64, kernel_size=3, padding=1),
            nn.BatchNorm2d(64),
            nn.LeakyReLU(0.1),
            SEBlock(64),
            nn.MaxPool2d(2),
            nn.Conv2d(64, 128, kernel_size=3, padding=1),
            nn.BatchNorm2d(128),
            nn.LeakyReLU(0.1),
            nn.AdaptiveAvgPool2d((7, 7))
        )
        self.classifier = nn.Sequential(
            nn.Linear(128*7*7, 512),
            nn.BatchNorm1d(512),
            nn.LeakyReLU(0.1),
            nn.Dropout(0.5),
            nn.Linear(512, num_class)
        )

    def forward(self, x):
        x = self.features(x)
        x = x.view(x.size(0), -1)
        x = self.classifier(x)
        return x
```

Fig. 6 Code Example of the CNN Model

The model consists of two primary modules: SEBlock (Channel Attention) module enhances main channel feature responses while suppressing less important ones through global average pooling, feature map compression, and fully connected layer processing to the convolutional model. The simplecnn (main model) module divides into feature extraction and classification parts. The feature extraction module consists of three convolutional layers. The first convolutional layer takes an input of 3 channels and outputs 32 channels, with a convolution kernel of 3×3 and a halved output size. The second convolutional layer takes an input of 32 channels and outputs 64 channels, with the channel attention influenced by the SEBlock training parameters and the convolution result size being halved again. The third convolutional layer takes an input of 64 channels and outputs 128 channels, with a uniform output size of 7×7. During the convolution process, the code nn.BatchNorm2d(32) is used to normalize the 32 channels, which provides accelerated training and alleviates the vanishing gradient problem. In the classification part, in addition to the conventional classification model call, we also use nn.Dropout(0.5) to invoke the random dropout model, randomly discarding neurons with a probability of 50%, and employing the method of randomly discarding neurons to prevent overfitting. The specific functions and parameter settings of each layer are detailed in Table 1.

Table 1. Meaning of Each Code and Specific Parameter Settings of the Model

Layer Type	Function	Key Parameters or Operation
Convolutional Layer (Conv2d)	Extracts local features and increases the number of channels	Kernel size, number of input/output channels, padding (to preserve size)
Batch Normalization (BatchNorm)	Normalizes output to accelerate training convergence	Learning mean and variance for each channel
LeakyReLU	Introduces non-linearity, allows activation of negative values	Negative slope (0.1)
Max Pooling (Maxpool2d)	Reduces computation and enhances translation invariance	Pooling window size (2×2)
SEBlock	Channel attention mechanism that dynamically adjusts channel weights	Global average pooling + fully connected layers for compression and restoration
Adaptive Average Pooling	Unifies feature map size to match fully connected layer input	Fixed output size (e.g., 7×7)
Fully Connected Layer	Maps feature representations to class space to perform classification	Input/output dimensions, activation function
Dropout	Randomly drops neurons to prevent overfitting	Dropout probability (0.5)

As film trailers directly affect promotional effectiveness and increasingly determine film evaluation through short video platform prosperity, we focus predictions on trailers. Good trailers bring audiences to movie scenes quickly, enabling emotional experience and plot curiosity that drives cinema attendance and box office contribution. The Convolutional Neural Network links warm and cool tone changes in trailer images with audience emotional fluctuations, serving as a reference variable for film box office prediction.

To achieve good predictive fit, we downloaded 200 film trailers from the dataset and captured frames every two seconds. We selected fifty trailers as the training set, manually inspecting and classifying images by distinguishing warm and cool tones while discarding distracting characteristics including cover images, actor introductions, plot subtitles, and black/white-dominated images to enhance training accuracy and reduce noise impact. Through model training, we obtained a model weight file with approximately 98.21% accuracy (see Fig. 7).

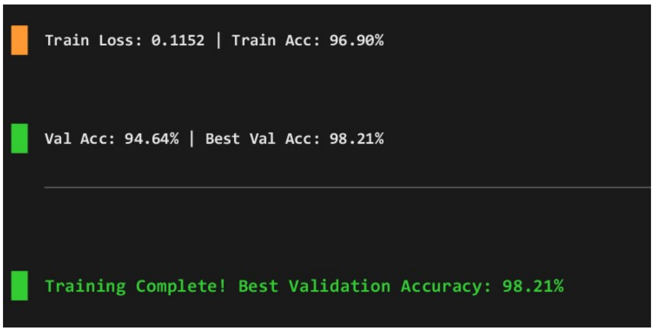

Fig. 7 Training Results of the Model Weight File

After training the model, we input the collected film trailer data into the model for emotional quantification analysis to obtain the total number of tone fluctuations in each film's trailer. Considering that the duration of trailers varies from film to film, we combined the number of tone fluctuations with the trailer's duration to derive a variable called "color change frequency." Before incorporating this variable into the predictive model, we conducted an emotional trend analysis of each film's trailer using the CNN, taking the film The Eight Hundred as an example (see Fig. 8 and Fig. 9).

From the figure, we can determine the emotional volatility rate of the film trailer and the average proportion of warm tone distribution in the trailer, which serves as one of the dependent variables in the model and is included as a feature.

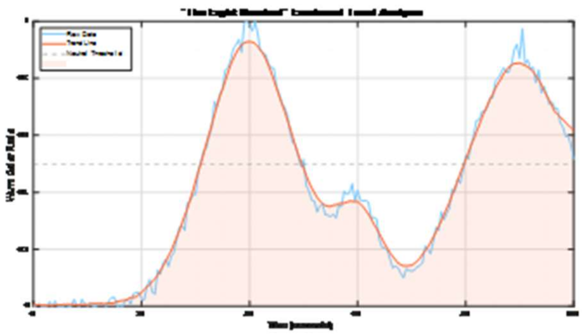

Fig. 8 Emotional Trend Analysis Curve of The Eight Hundred

Fig. 9 Emotional Analysis Results of The Eight Hundred

3.4. Establishment of the Random Forest Model

Random Forest (RF) is a machine learning algorithm [5] that is an ensemble of decision trees (CART). Each decision tree within the ensemble operates independently and can handle both classification and regression problems. For a given input sample, each tree produces a classification or prediction result. The Random Forest model aggregates all the classification results (or averages the regression predictions) and designates the class with the highest number of votes as the final output.

Sampling with replacement is conducted from the original dataset to construct sub-datasets, with the data volume of each sub-dataset being identical to that of the original dataset. Elements can be repeated both across different sub-datasets and within the same sub-dataset. Similar to the random selection of the dataset, not all candidate features are used in each split process of the subtrees in the Random Forest. Instead, a certain number of features are randomly selected from all the candidate features, and then the optimal feature is chosen from these randomly selected features. The decision tree process of the Random Forest is shown in Fig. 10:

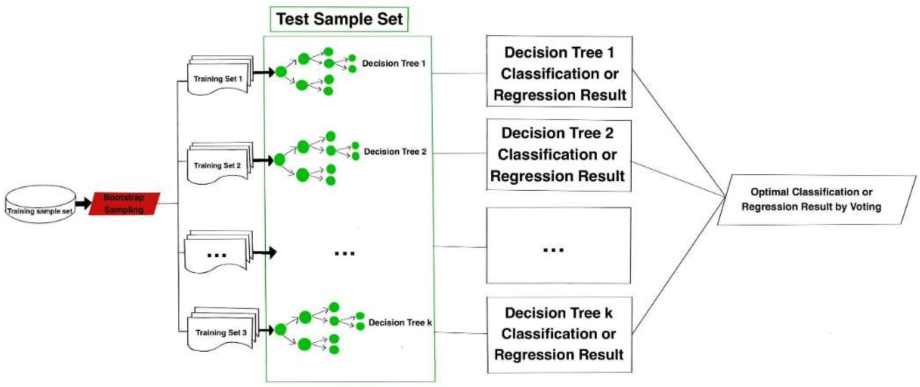

Fig. 10 Random Forest Decision Tree Model

The Random Forest algorithm was implemented using MATLAB. The emotion values quantified by CNN, the actor suitability established by the heatmap matrix, and the collected film genres, release years, film durations, subscription numbers, whether the film is in a peak season, and whether it is a series were used as multiple independent variables in the model. The predicted box office of the film was used as the dependent variable in the model. The data was randomly shuffled, with 180 data points used as the training set and the remaining 40 data points used as the test set to verify the accuracy of the model, as shown in Fig. 11, 12, and 13.

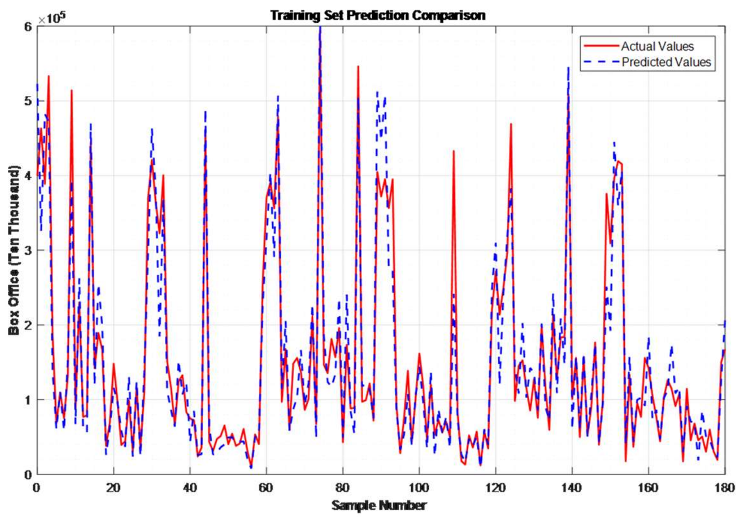

Fig. 11 Comparison of Prediction Results on the Training Set

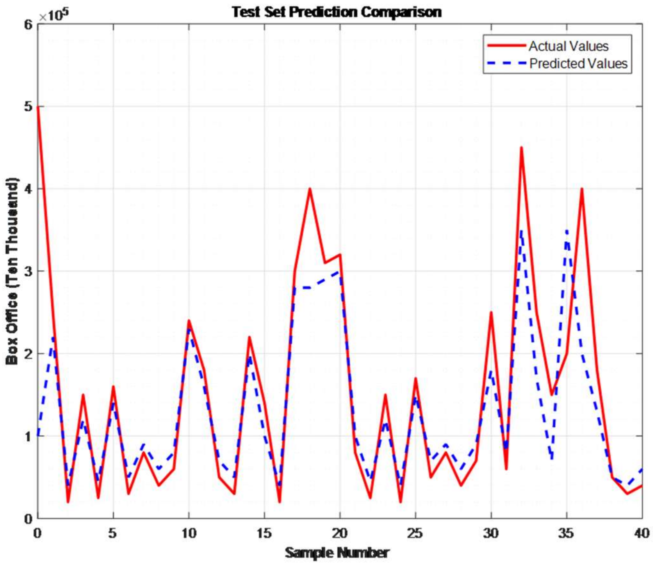

Fig. 12 Comparison of Prediction Results on the Test Set

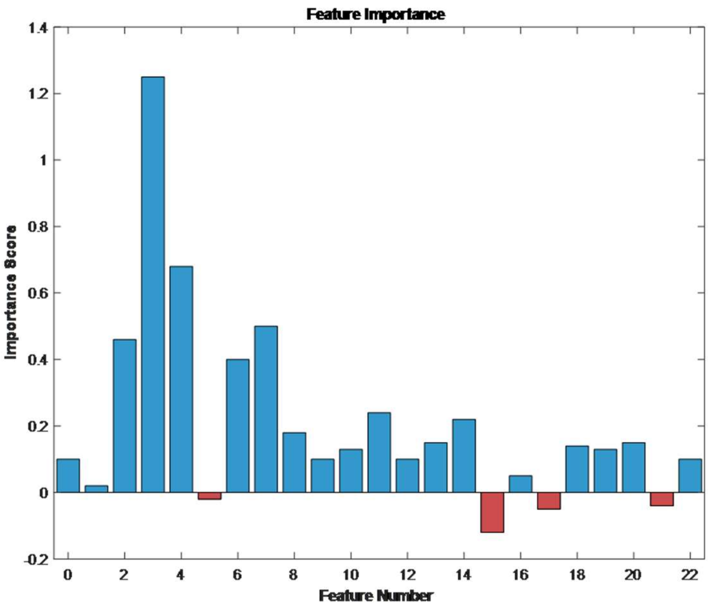

Fig. 13 Histogram of Variable Importance

3.5. Model Features

Research Findings: Through the Random Forest model, we identified five factors with greater positive impact on box office performance: the number of people wanting to see the movie, Maoyan rating, emotional volatility rate of the trailer, holiday release timing, and popular actors with good role fit. Additionally, whether a film is a comedy or a sequel is an important characteristic of high box office movies. However, popular directors, release year, and film duration do not significantly impact box office performance. Using the Random Forest model, we can forecast final box office of newly released movies, scientifically guiding film production quality upgrades to promote a healthy industry ecosystem and better implement cultural confidence strategy through film.

Model Analysis (Advantages): 1. Random Forest can analyze data with very high dimensionality (many features) without the need for dimensionality reduction or feature selection. 2. It can determine the importance of each feature. 3. It can assess the interactions between different features. 4. It is less prone to overfitting compared to BP neural networks, which often suffer from poor prediction accuracy due to overfitting. 5. The training speed is relatively fast and it can be easily parallelized. 6. It is relatively simple to implement. 7. It can balance errors for imbalanced datasets. 8. It can maintain accuracy even if a large portion of the features are missing [5].

Model Analysis (Disadvantages): 1.Random Forest is more complex and computationally expensive compared to decision tree algorithms. 2. Due to its inherent complexity, it requires more time to train compared to other similar algorithms.

3.6. Error Analysis

Absolute Error (Absolute Error, AE)

Formula:

$$AE = | y_{\text{actual}} - y_{\text{predicted}} | \tag{1}$$

Significance: It directly measures the absolute value of deviation between predicted values and actual values.

Relative Error (Relative Error, RE)

Formula:

$$RE = \frac{AE}{y_{\text{actual}}} \times 100\% \tag{2}$$

Significance: Normalized error reflects the proportion of the error relative to the actual value.

Mean Absolute Error (Mean Absolute Error, MAE)

Formula:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_{\text{actual},i} - y_{\text{predicted},i} | \tag{3}$$

Significance: It measures the overall prediction accuracy of the model; the smaller the value, the better.

Calculation Results:

$$MAE = \frac{32,272 + 60,260 + 14,669}{3} = 357.34 \text{ million} \tag{4}$$

Root Mean Squared Error (Root Mean Squared Error, RMSE)

Formula:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{\text{actual},i} - y_{\text{predicted},i})^2} \tag{5}$$

Significance: It is more sensitive to larger errors, reflecting the stability of predictions.

Calculation Results:

$$RMSE = \sqrt{\frac{(32,272)^2 + (60,260)^2 + (14,669)^2}{3}} \approx 408.53 \text{ million} \tag{6}$$

Mean Absolute Percentage Error (Mean Absolute Percentage Error, MAPE)
Formula:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_{\text{actual},i} - y_{\text{predicted},i}}{y_{\text{actual},i}} \right| \tag{7}$$

Calculation Results:

$$MAPE = \frac{3.506\% + 16.771\% + 11.91\%}{3} \approx 10.73\% \tag{8}$$

Significance: A MAPE greater than 12% indicates that the model performs well in terms of relative error.

3.7. Model Validation

Table 2. Model Validation Results for 2025 Films

Film Released in 2025	Predicted Final Total Box Office (Unit: 10,000)	Current Box Office (as of April 26)	Relative Error (%)
Octopus with Broken Arms	90039	93311	3.506
Detective Chinatown 1900	299040	359300	16.771
Creation of the Gods II	137769	123100	11.91

Evaluation:
As shown in Table 2, this study conducted an error analysis on the box office predictions of three films scheduled for release in 2025. The results revealed significant heterogeneity in the model's predictive performance. Specifically, the prediction for Octopus with Broken Arms exhibited a high degree of consistency with the actual observed values, demonstrating superior prediction accuracy compared to other samples. In contrast, Detective Chinatown 1900 showed the greatest prediction deviation, with a relative error reaching 16.8%. It is worth noting that the relative errors of all predicted samples were strictly controlled below the critical value of 20%, with an overall average absolute percentage error of 10.7%, in accordance with the following formula:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_{\text{actual}}^{(i)} - y_{\text{predicted}}^{(i)}}{y_{\text{actual}}^{(i)}} \right| \tag{9}$$

This quantitative indicator suggests that the model possesses robust box office prediction performance.

4. Research Conclusion

4.1. As an Evaluation System for Film Trailers

With the continuous advancement of the information age and the improvement of economic standards, film has evolved from pure artistic entertainment into a sophisticated vehicle for ideological and cultural transmission, as exemplified by The Shawshank Redemption and The Wandering Earth, which demonstrate cinema's capacity to convey profound values and worldviews through narrative frameworks that resonate across cultural boundaries. While existing research on Chinese film box office prediction remains limited, previous models developed by Marshall et al. utilizing historical data with linear regression approaches, alongside Sawhney et al.'s cumulative viewer forecasting methodologies, prove inadequate for analyzing complex and diversified markets characterized by rapidly evolving audience preferences and multifaceted emotional engagement patterns that significantly influence box office performance. This study addresses these methodological gaps by adopting a multi-model combination approach that integrates time series analysis for industry trend identification, heatmap matrix models for assessing actor-genre compatibility based on the recognition that effective casting generates substantial celebrity effects, and Random Forest ensemble learning that demonstrates robust predictive capability through achieving satisfactory validation results for three films scheduled for 2025 release with relative errors controlled below 20%. Beyond conventional financial forecasting, the proposed framework enables systematic analysis of films' embedded values, worldviews, and emotional attitudes, thereby facilitating predictions of audience acceptance patterns while supporting the propagation of culturally significant narratives and promoting global dissemination of excellent traditional Chinese culture through accessible and visually impactful cinematic means, ultimately establishing a dual-purpose system that functions simultaneously as both a commercial prediction tool and a comprehensive cultural evaluation mechanism.

4.2. Promoting Cultural Confidence through the Film Industry

Film functions as a sophisticated modern translator that converts China's millennia-old civilization into emotional resonance through contemporary audiovisual language, with recent successful productions such as Ne Zha's reinterpretation of traditional mythology through modern narrative techniques and Chang An's visualization of classical poetic imagery effectively activating the genetic code of traditional culture while constructing spiritual bridges connecting historical heritage with contemporary expression. In the realm of international cultural communication, Chinese cinema has transcended the mere display of cultural symbols to enter an era of substantive value dialogue, exemplified by The Wandering Earth series' replacement of Western individual heroism narratives with shared future concepts rooted in collective responsibility, alongside The Creation of the Gods trilogy's employment of advanced digital technology to represent Bronze Civilization aesthetics with unprecedented visual fidelity, thereby signaling Chinese cinema's qualitative leap from "telling Chinese stories well" to "building independent discourse systems" capable of engaging global audiences on equal footing. The establishment of scientific box office prediction models serves strategic cultural objectives beyond commercial purposes by providing data-driven insights into audience preferences and market trends that guide the film industry toward maintaining cultural resilience while pursuing commercial success, enabling Chinese cinema to achieve a fundamental transformation from industrial output focused on quantitative metrics to civilizational dialogue emphasizing qualitative cultural depth and authentic value expression that strengthens China's cultural confidence and expands its soft power influence within the increasingly interconnected global cultural marketplace.

References

[1] X. Hu, Y. Guo, M. Feng, et al., "Time Series Model Prediction and Analysis of Malaria Cases in Jiangsu Province from 1980 to 2023," Chinese Journal of Frontier Health and Quarantine, vol. 47, no. 6, pp. 582-587, 2024.

[2] C. Deng, J. Song, R. Sun, et al., "Visual Analysis System for Cigarette Market Data Based on Heatmap," Tobacco Science & Technology, vol. 49, no. 12, pp. 91-97, 2016.

[3] C. Yan and C. Wang, "Development and Application of Convolutional Neural Network Models," Journal of Frontiers of Computer Science and Technology, vol. 15, no. 1, pp. 27-46, 2021.

[4] X. Ding, "Research on the Application of BP Neural Networks and Convolutional Neural Networks in Text Recognition," Ph.D. dissertation, Huazhong University of Science and Technology, Wuhan, China, 2017.

[5] K. Fang, J. Wu, J. Zhu, et al., "A Review of Random Forest Methods," Statistics & Information Forum, vol. 26, no. 3, pp. 32-38, 2011.