International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n2.29
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Analysis and Box Office Prediction of Chinese High Box Office Films- Empirical Analysis Based on the Top250 Movies on Maoyan.com

Qinrui Hu *

Wenzhou Kean University, Wenzhou, Zhejiang, 325000, China
*Corresponding Author: Qinrui Hu (Email: huqinr@kean.edu)

ABSTRACT

This project focuses on analyzing and predicting box-office performance using data from Maoyan.com. Several key factors, including release year and month, ticket price, movie duration, and the proportion of positive reviews, are examined for their influence on box-office revenue. We begin by employing simple linear regression to investigate the relationship between Maoyan.com ratings for the top 250 movies and their box-office performance. Additionally, we construct an Autoregressive Distributed Lag (ARDL) model to assess the impact of these factors over time. The model parameters are estimated using the least squares regression method, and the Mean Absolute Percentage Error (MAPE) is minimized to achieve optimal prediction accuracy. Extensive data cleaning and preprocessing ensure the reliability of our analysis. The findings from this study provide valuable insights into the determinants of box-office success and offer a robust predictive model for future revenue estimation.

KEYWORDS

Box-Office Performance Prediction; Maoyan.com; Linear Regression; Time series analysis; Mean Absolute Percentage Error (MAPE)

1. INTRODUCTION

1.1. Background

The rapid evolution of the film industry has dramatically altered how movies are produced, marketed, and consumed. In today's digital era, several key factors such as movie genre, director, actors, plot summary, and marketing activities play a crucial role in determining a film's box office performance (Zhang et al., 2009; Lipizzi et al., 2016). Previous research has consistently shown that these factors collectively influence consumer purchase decisions and, consequently, box office revenues (Ding et al., 2017; Hur et al., 2016) [1].

Despite extensive research on box office prediction, there remains a significant gap in understanding the nuanced impact of various features, especially those derived from online review platforms. The growing influence of online reviews and ratings makes it essential to explore their role in predicting box office performance. This study aims to bridge this gap by leveraging data from Maoyan.com, a popular Chinese movie review and rating platform, to analyze the top 250 movies.

1.2. Objectives and Scope

The primary objectives of this study are threefold:

Analyze Impact Factors: To analyze the impact of average ticket price, average attendance, release year, and release month on box office revenue.

Investigate Relationships: To investigate the relationship between ticket price and average attendance on Maoyan.com and box office performance using simple linear regression.

Develop Predictive Models: To develop and evaluate predictive models using Autoregressive Distributed Lag (ARDL) models and other regression techniques to forecast box office revenues based on identified factors.

This study focuses on factors influencing box office revenues of the top 250 movies listed on Maoyan.com, specifically examining average ticket price, average attendance, positive review proportion, and release date. Various predictive modeling techniques, including Linear Regression, KNN, Random Forest, and ARDL models, will be employed. The study is geographically focused on China, with model performance evaluated using Mean Absolute Percentage Error (MAPE). Limitations include the data scope from Maoyan.com and potential unaccounted future trends.

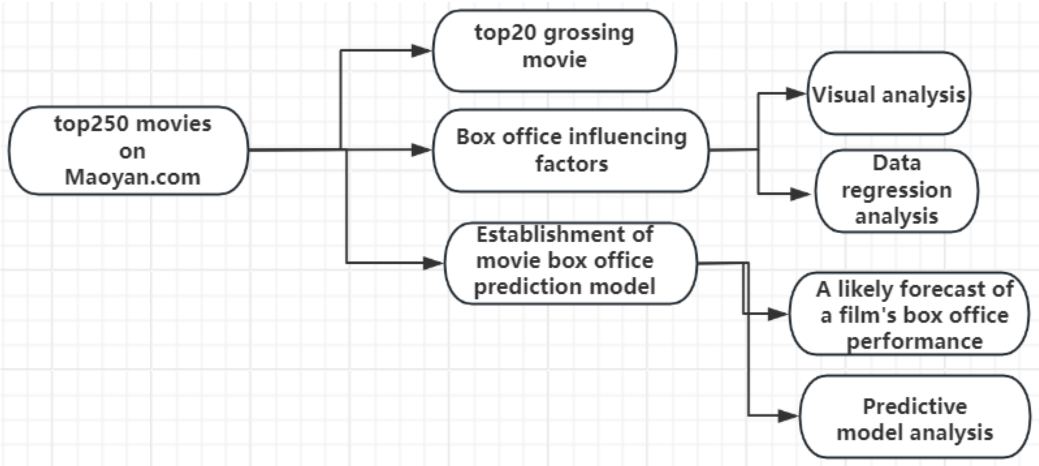


Figure 1. Factors Influencing Box Office Performance

1.3. Literature Review

Several studies have explored the relationship between online reviews and box office performance. Liu (2006) found that the volume of reviews significantly impacts box office revenue. Mishne and Glance (2006) demonstrated that review volume from social media platforms positively correlates with box office earnings. Dellarocas et al. (2007) highlighted that both the quantity and valence (positive or negative) of reviews affect box office outcomes. Duan et al. (2008) further showed that movie reviews and box office revenues mutually influence each other.

However, many of these studies do not fully consider the dynamic nature of consumer preferences and the temporal aspects of box office performance. The integration of time series analysis with predictive modeling, as suggested by Du, Xu, and Huang (2010), remains an area ripe for further exploration. This study aims to build on these insights by incorporating both traditional regression techniques and advanced machine learning models to predict box office revenues. Using data from Maoyan.com, this research will specifically address the impact of average ticket price, average attendance, positive review proportion, and release date on box office performance. Additionally, this study will explore the effectiveness of different lag structures in time series data, an area less extensively studied in the existing literature.

1.4. Problem Statement

The primary research question addressed in this study is: How do release timing, ticket price, movie duration, and positive review proportion affect box office revenue? To answer this question, a systematic approach will be followed, including data collection, preprocessing, exploratory data analysis, model fitting, and evaluation. This study aims to provide comprehensive insights into the factors that significantly impact box office performance. By integrating various statistical and machine learning techniques, we will develop a robust predictive model to understand and forecast box office revenue. The findings are expected to guide filmmakers and marketers in making informed decisions to enhance their films' success.

2. ANALYSIS (HIGH BOX OFFICE FACTOR)

To better understand and analyze the factors influencing movie box office performance, I first conducted data visualization. Figure 2 shows the top 20 grossing movies and their box office revenues. It can be seen that the box office revenues of these movies range from 3.1 billion to 5.7 billion.

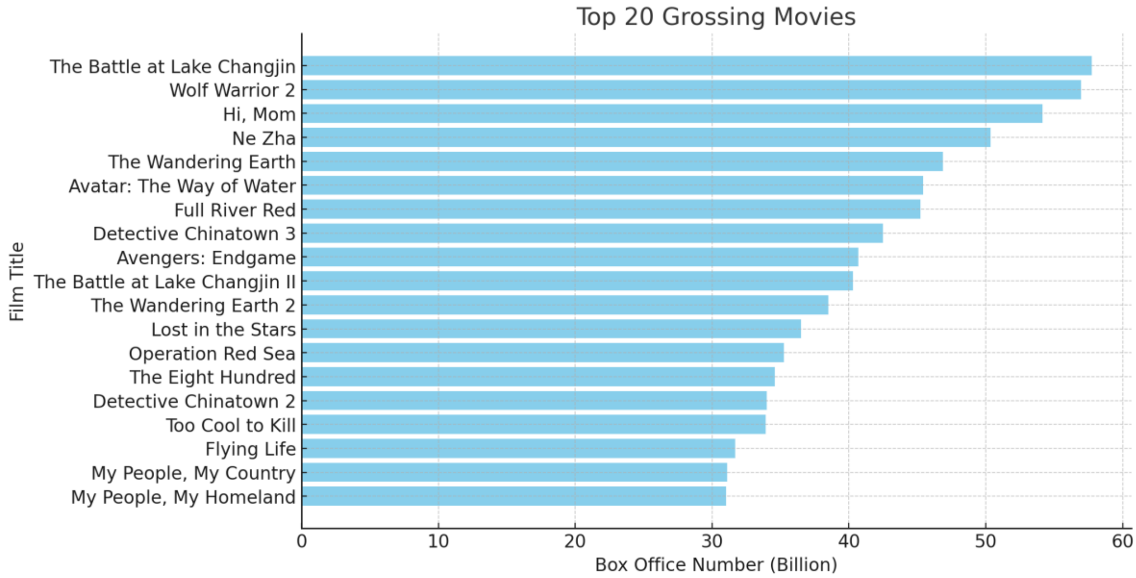
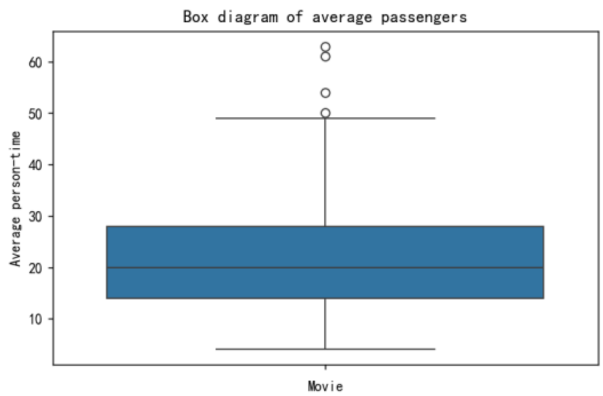


Figure 2. Top 20 Grossing Movies

Next, I investigated the relationship between average ticket price and average person-time, which reflects box office volume. Figures 3 and 4 display the box plots of average person-time and average ticket price, respectively. It can be seen that both average ticket price and average person-time have some outliers. The scatter plots Figure 5 and 6 show their relationship with the year. Clearly, as the years progress, the average person-time decreases while the average ticket price increases. This indicates that recent movies have become more expensive, and fewer people are attending each screening. This trend also reflects some issues in the Chinese film industry, such as "high ticket prices" and "phantom theaters" manipulating box office figures.

These visualizations provide a foundation for subsequent regression analysis. By initially observing the data distribution and trends, we can better understand the potential impact of different factors on box office revenue, which informs model construction and parameter estimation.



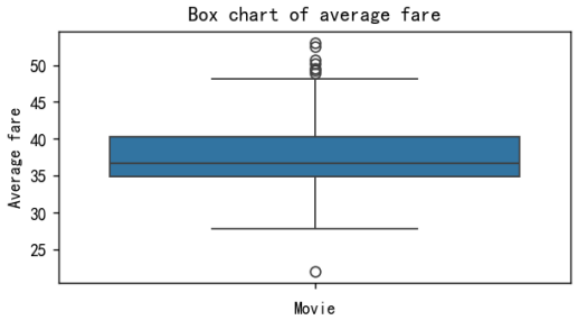


Figure 3. Box diagram of average passengers **Figures 4.** Box chart of average fare

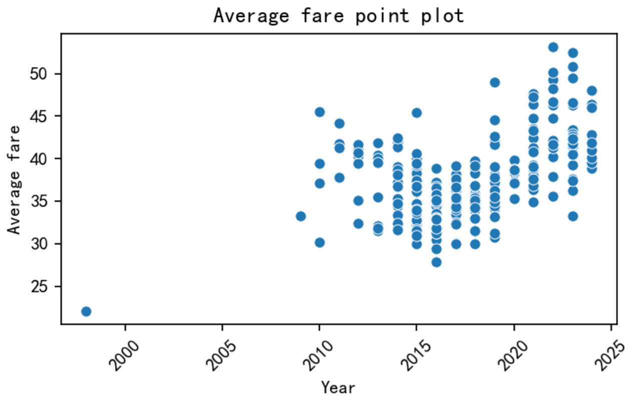
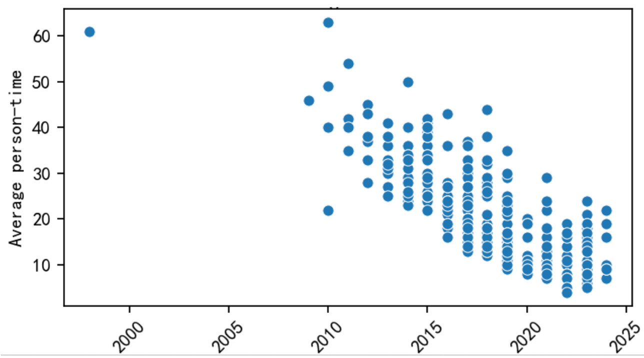


Figure 5. Scatter plot showing relationship between average person-time and year



Figures 6. Scatter plot showing relationship between average ticket price and year

Figure7 illustrates the annual distribution of top-grossing films from 1998 to 2024. A significant increase starts around 2013, peaking in 2016, 2017, and 2019 with 31 high-grossing films each year. This trend suggests substantial growth in the Chinese film industry over the past decade. The decline in 2020 reflects the impact of the COVID-19 pandemic, with a subsequent recovery in 2021 and 2022. Recent fluctuations indicate ongoing adjustments in the post-pandemic era.

Figure8 shows that high-grossing films are commonly released in January, February, and December. These months correspond to the Chinese New Year and end-of-year holiday seasons, times when audiences are more available and willing to spend. In contrast, August and November have fewer high-grossing releases, likely due to less favorable market conditions. This seasonal trend highlights the importance of strategic release timing for maximizing box office performance.

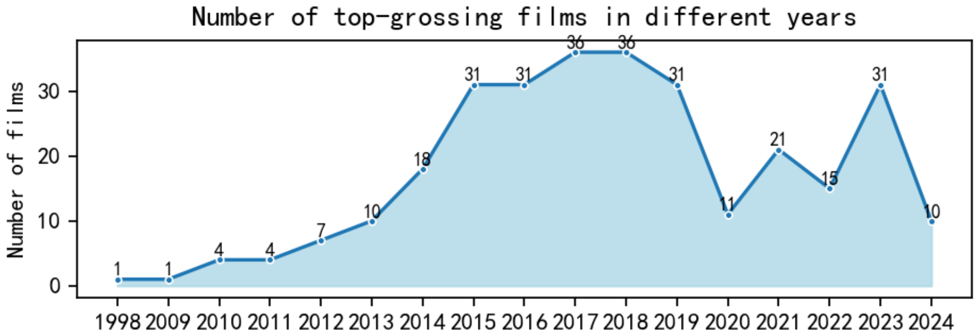


Figure 7. Number of top-grossing films in different years

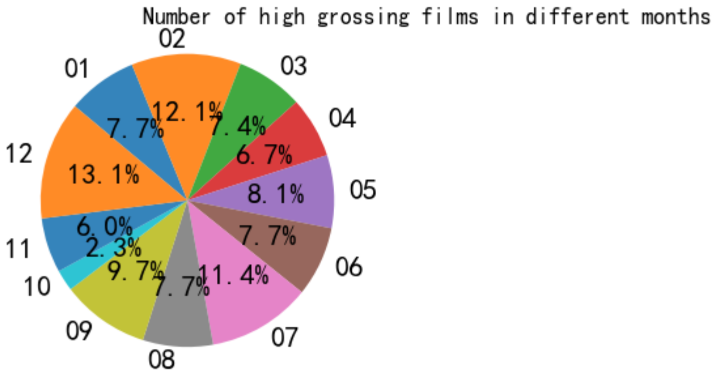


Figure 8. Number of high grossing films in different months

3. METHODOLOGY

3.1. Model Construction and Evaluation

An ARDL model is constructed to study the impact of early box-office revenue, film duration, and ticket price on overall box-office revenue. The model parameters are adjusted to minimize the Mean Absolute Percentage Error (MAPE), ensuring the highest prediction accuracy. To study the impact of early box office revenue, film duration, percentage of praise， and ticket price on the overall box office revenue during the release period, a box office prediction model is constructed. This model is fitted using an Autoregressive Distributed Lag (ARDL) model [4], as represented in equation (1).

$$y_t = \varphi_0 + \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{j=1}^{k} \sum_{i=1}^{q} \beta_{ij} X_{i,t-j} + \mu_t \tag{1}$$

The model includes the dependent variable with p lags, k exogenous variables, and q_i lags for each exogenous variable. The lag order for each exogenous variable is q_1, q_2, and so on. p represents the number of days for early box office revenue, k is the number of exogenous variables, and q_i is the number of days for each exogenous variable. The parameters $\varphi_i$and β_i are estimated based on the research data. In this study, the exogenous variables are film duration, percentage of praise and ticket price, so k=3 is set. The initial values for p, q_1, and q_2 are set to 7, q1and q2, respectively. The sample data is tested continuously, adjusting these parameters to minimize the Mean Absolute Percentage Error (MAPE). The parameter values with the optimal prediction accuracy are selected as the final values. The calculation of MAPE is shown in equation (2).

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Predict_i - True_i}{True_i} \right| \tag{2}$$

n represents the total number of predictions, $Predict_i$ represents the predicted box office revenue, and $True_i$ represents the actual box office revenue. The smaller the MAPE value, the better the prediction performance of the model.

3.2. Stationarity Test

Taking the box office revenue time series [3] of the movie "Wolf Warrior 2" as an example, the line chart is presented in Figure 9. The KPSS test results for the processed time series are shown in Table 6. The test statistic value is compared with the critical values at the 1%, 5%, and 10% significance levels. Since the test statistic value is below the critical values, we fail to reject the null hypothesis and conclude that the time series is stationary. Extensive testing on the sample data consistently demonstrates that the series passes the stationarity test using the KPSS method.

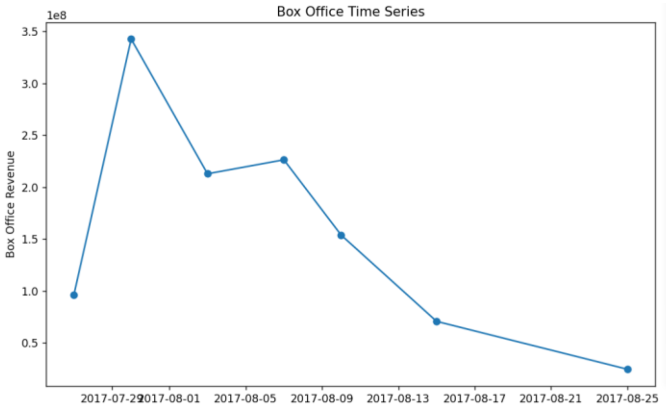


Figure 9. Box Office Time Series

```
KPSS Statistic: 0.34412155974229064
p-value: 0.1
Critical Values: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
The time series is stationary according to the KPSS test (p >= 0.05).

Process finished with exit code 0
```

3.3. Final Model Establishment and Model Evaluation

Using the KNN algorithm, we input the time series data for box office revenue, positive review proportion, and release time—each having passed the stationarity test—into our pre-constructed model. We then employ the least squares regression method to estimate the parameters and predict box office revenue. By continuously adjusting the lag orders of the box office revenue, positive review proportion, and release time, we strive to find the combination of lag orders that minimizes the Mean Absolute Percentage Error (MAPE). This iterative process leads to the development of the final box office prediction model. To validate the model's performance and ensure it performs well on unseen data, we split the dataset into training and testing sets and used cross validation to evaluate the model's performance. By comparing the MAPE values of the KNN and linear regression models, we found that the linear regression model slightly outperformed the KNN model, with a MAPE value of 0.6423 compared to 0.6519 for the KNN model. This indicates that the linear regression model has better predictive accuracy. While adjusting the lag orders, we continuously tried different combinations and found that the optimal lag order was (4, 1, 2), resulting in the final model:

$$y_t = \varphi_0 + \sum_{i=1}^{4} \varphi_i y_{t-i} + \sum_{j=1}^{1} \sum_{i=1}^{2} \beta_{ij} X_{i,t-j} + \mu_t \tag{3}$$

4. RESULT AND DISCUSSION

4.1. Model Summary

The regression model, which incorporates factors such as average ticket price and average person-time, has an R-squared value of 0.186. This indicates that these variables explain approximately 18.6% of the variance in box office revenues for the top 250 films listed on Maoyan.com. The result is not surprising, given that the length of the ticket and the small difference in price usually do not influence the purchase decision of consumers. While this figure suggests a moderate explanatory power, it highlights the complexity and multifactorial nature of box office success, suggesting that other unmodeled factors may also play significant roles. The model's F-statistic of 33.68 with a highly significant P-value (6.71e-14) underscores the overall statistical significance of the predictors included. Specifically, the coefficients for average ticket price (0.8876) and average person-time (0.2002) are both significant at the $p < 0.0001$ level, indicating a strong positive relationship with box office revenue. Despite the significant findings, the model diagnostics indicate potential issues that could affect the interpretation and reliability of the results. And the regression analysis results indicate a significant positive correlation between box office revenue and ratings, with an R^2 value of 0.916, meaning that box office revenue explains 91.6% of the variation in ratings. The regression coefficient for box office revenue is 0.0010 and is highly significant (p-value much less than 0.05). Additionally, the residuals are approximately normally distributed (Omnibus test p-value of 0.286, Jarque-Bera test p-value of 0.518) and show no significant autocorrelation (Durbin-Watson statistic of 2.409). This means that ratings play a big role in how consumers buy tickets. Imagine if you want to know what's going to be a good movie recently. We often use ratings to determine our viewing choices.

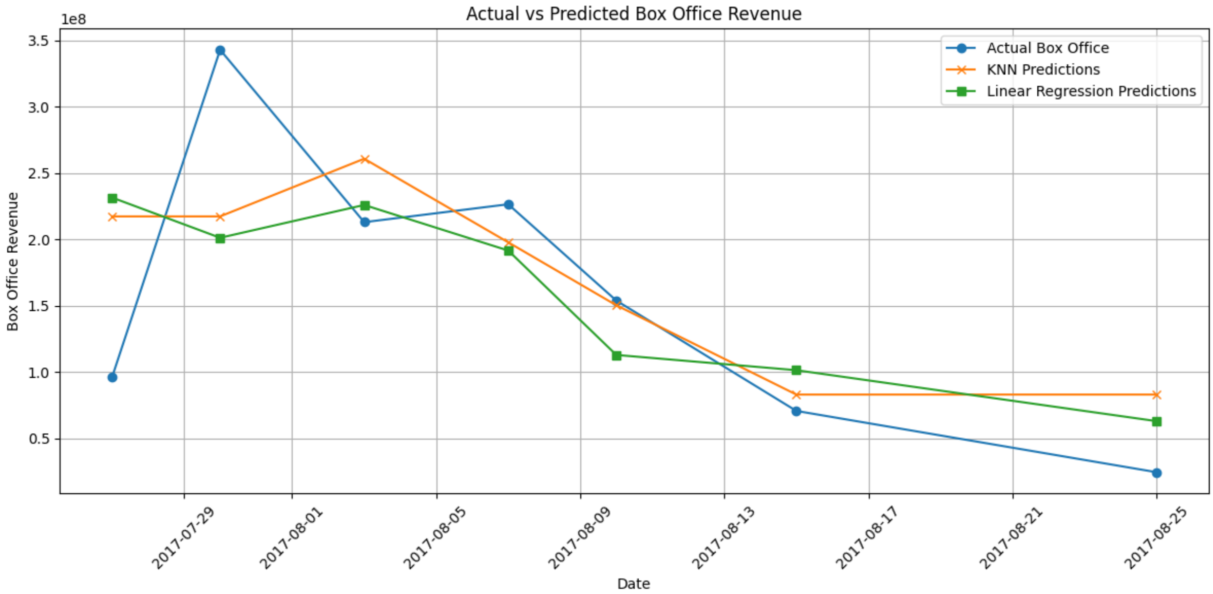


Figure 10. Actual vs Predicted Box Office Revenue

The predictive modeling efforts, particularly the ARDL model and the machine learning approaches like KNN and linear regression, have demonstrated varied success. The best MAPE achieved was significantly low, which showcases the potential of our modeling approach to make accurate predictions. However, the complexity of choosing appropriate lag structures and balancing between different predictors' influence remains a challenging aspect of forecasting box office revenues.

The insights derived from this analysis are intended to assist filmmakers and marketers in strategizing more effectively to enhance a film's market success. By understanding the influence of ticket pricing, review dynamics, and timing of the release, stakeholders can make informed decisions to maximize returns.

4.2. Short Comings

The KNN and linear regression models have their limitations. The KNN algorithm, while flexible, may struggle with high-dimensional data and can be sensitive to the choice of k and the distance metric. Additionally, KNN does not inherently provide insight into the relationships between variables, which can be a drawback for interpretability. The linear regression model, although easy to implement and interpret, assumes a linear relationship between predictors and the outcome, which may not always be the case. It also requires that the predictors are not highly collinear and that the residuals are normally distributed and uncorrelated, which, as seen in the diagnostics, may not hold true for our data.

4.3. Future and Prospect

Future research may benefit from expanding the set of explanatory variables to include additional elements such as social media engagement metrics, broader demographic data on viewers, and macroeconomic factors. Additionally, addressing the diagnostic concerns within the model could improve its accuracy and reliability. Implementing advanced machine learning techniques that can handle non-linear relationships and interaction effects may also provide deeper insights into the dynamics of box office revenues.

In conclusion, while this study has made substantial contributions to understanding the determinants of box office success in the Chinese market, the ongoing evolution of the film industry and viewer behaviors will necessitate continual updates and refinements to predictive models employed within this domain.

REFERENCES

[1] He, Q., & Hu, B. (2021). Research on the Influencing Factors of Film Consumption and Box Office Forecast in the Digital Era: Based on the Perspective of Machine Learning and Model Integration. Wireless Communications & Mobile Computing (Online), 2021, https://doi.org/10.1155/2021/6094924

[2] Ya-Han, H., Wen-Ming Shiau, Sheng-Pao Shih, & Cho-Ju, C. (2018). Considering online consumer reviews to predict movie box-office performance between the years 2009 and 2014 in the US. The Electronic Library, 37(6), 1010-1026. https://doi.org/10.1108/EL-02-2018-0040

[3] Muzammil, H. S., & Muhammad, A. I. (2023/11/09/). Investigation of time series-based genre popularity features for box office success prediction. PeerJ Computer Science, https://doi.org/10.7717/peerj-cs.1603

[4] ZHU Rui, MA Yong-mei, CHEN Ya-nan. Research on Online Comment in Box Office Forecasting: Empirical Analysis Based on Chinese Mainland Film Market [J]. Journal of Chaohu University, 2021, 23(5): 71-79.