Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/116/20251729

A Review of Early Box Office Prediction Based on Social Media

Fangrui Liu

International College, Chongqing University of Posts and Telecommunications, No. 2, Chongwen Road, Nan'an District, Chongqing, China

2842446520@qq.com

Abstract. A wide range of box office prediction surveys based on artificial intelligence have made progress. However, early box office prediction remains imperfect due to significant uncertainty before a film's release, leading to increased interest in this area of research. Despite this, research is still in its early stages. To address this issue, we review early box office prediction methods based on social media data. By examining three approaches, we derive a generic process for early box office prediction. This paper compares the advantages and disadvantages of machine learning- and deep learning-based prediction methods. Additionally, we analyze the research trends and challenges in the field, providing a reference for researchers who are new to this area.

Keywords: Box Office Prediction, Machine Learning, Deep Learning.

1. Introduction

Media can be used to analyze public mood and opinion, which is beneficial for sentiment analysis and model predictions. By collecting data from social media, a variety of features can be extracted to build prediction models. More successful movies benefit from accurate predictions, leading to favorable market outcomes.

However, early box office prediction is imperfect due to the high level of uncertainty before a film's release, which has prompted increased research in this area. For example, access to real-time information, and whether the information is official or authoritative, vary across platforms. Additionally, data published on these platforms may be processed or altered and may not always align with actual facts. These challenges persist, and research in this field is still in its early stages. Many existing prediction methods have significant errors and limitations that require further improvement.

To address these issues, we review previous research conducted since 2019. Numerous factors influence box office performance, such as the number of films released simultaneously, the style and scope of advertising, and whether the release coincides with a special holiday. Identifying features that influence a film's success, whether positively or negatively, is crucial. In this paper, we compare the research methods and prediction results from various studies to synthesize an improved prediction approach. Even without access to real-time social media data, many methodologies can still be applied. We review machine learning methods, including stacking model fusion, random forest, k-nearest neighbor algorithm, extreme gradient boosting, and light gradient boosting machine. This review

DOI: 10.54254/2755-2721/116/20251729

provides valuable references and data for scholars who are new to this field and building prediction models.

2. Literature Review

An overview of current early box office prediction methodologies based on social media can be broken down into four main steps, as shown in Figure 1. The process begins with data input, followed by feature extraction and model building, and concludes with model assessment. Based on the artificial intelligence algorithms employed, these methods can be categorized into two types: early box office prediction methods using classical machine learning and those using deep learning.

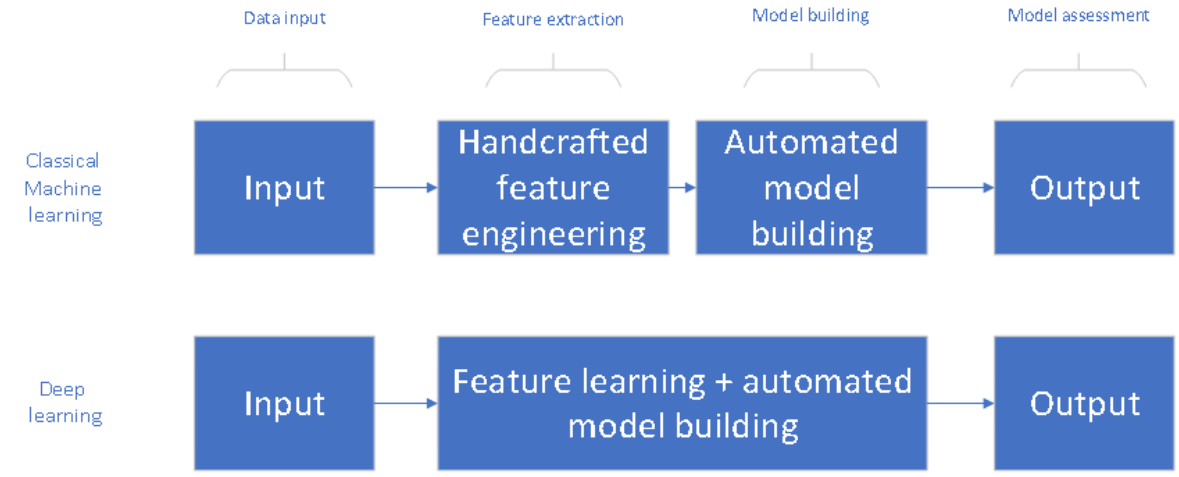

Figure 1. Process of 2 type analytical model building.

2.1. Classical Machine Learning-based Method

The earliest researchers used machine learning methods, such as decision trees, for early box office prediction. For instance, Sudais Muhammad et al. found that among various machine learning methods, the decision tree methodology outperformed others in classification [1]. Additionally, the ensemble technique enhanced their model's ability to classify datasets more effectively. However, one limitation was that the results were difficult to interpret. Later, other researchers focused on different ensemble learning methods. For example, Liao Yi et al. proposed a stacking fusion model for box office prediction [2]. By analyzing the characteristics of movies, they employed an ensemble of machine learning algorithms, such as random forest and k-nearest neighbor. However, due to data limitations, they suggested collaboration with official studios to access internal data for more accurate predictions. Building on these methods, new scholars have further refined them. For instance, Ni Yuan et al. found that the enhanced stacking algorithm outperformed other methods like the light gradient boosting machine [3]. However, their study was limited by the small dataset used for box office prediction. Other researchers have continued studying these issues. For example, Lu Wei et al. developed a core index system that influences the managed business object, which could predict outcomes using the analytic hierarchy process [4]. However, the model's accuracy and completeness need improvement, which is associated with the availability of public data. On the other hand, some researchers have conducted studies based on platform data sources. For instance, Song Tingting et al. found that customers access services and consume content from microblogging platforms, similar to third-party platforms [5]. Building on previous surveys, some scholars have improved decision tree methods from a classification perspective. For instance, Wei et al. introduced an ensemble learning model using the Pearson correlation coefficient for feature selection [4]. They found that their model outperformed the decision tree methodology. Additionally, classifying movies according to specific ratings and building predictive models for each category enhances prediction accuracy. Some scholars have also worked on improving the speed of box office predictions. For instance, Masrury Riefvan Achmad et al. found that the Naïve

DOI: 10.54254/2755-2721/116/20251729

Bayes model computes faster than other models due to its lower computational requirements [6]. However, machine learning methods vary in terms of accuracy and speed. To address potential issues such as data overflow, some scholars have explored binary structures to refine their models. For example, Park Junghoon et al. trained a stacking model with two levels—zero and one—and employed diverse hyperparameter tuning, which improved their model's performance compared to using a single model [7]. It is also important to collect more diverse samples at each level, as this can enhance the performance of stacking. Expanding the dataset range is crucial for better results. Other researchers have explored different methodologies. For instance, Wang Dongqi et al. found that the random forest model outperformed other machine learning methods due to its lower mean absolute error [8]. However, the accuracy of their model was influenced by the audience size, especially when the audience was small. This was followed by additional research focused on the random forest model. For instance, Menaga D. et al. found that their random forest model performed better than other models, making it the most precise [9]. In addition to domain-specific features, some scholars have refined their exploration of non-domain-specific features. For example, Al Fahoum Amjed et al. optimized non-domain-specific features, making the k-nearest neighbor algorithm superior to other machine learning algorithms [10]. Notably, their model was significantly faster than others, particularly in fine-grained classification tasks.

2.2. Deep Learning-based Method

Deep learning has made significant breakthroughs in various research areas, including early box office prediction. Researchers have applied deep learning techniques to predict audience behavior more effectively. For instance, Ko Ming Ya et al. found that minimal intra-genre multimodal embeddings yielded more accurate predictions compared to other methodologies [11]. However, expanding datasets to include a broader range of viewers—such as different age groups, cultures, and preferences—would enhance prediction accuracy. Additionally, developing methods to capture facial expression changes corresponding to specific content during viewing could further improve predictions. In addition to using single features for forecasting, more researchers are adopting multiple features for early box office prediction. For example, Madongo Canaan Tinotenda et al. proposed using multimodal features, finding that complex features could serve as advanced inputs for neural network algorithms [12]. However, they overlooked the potential of incorporating sentiment analysis using natural language processing. Subsequent researchers have refined this approach. Sindhu Irum et al. analyzed a broad range of social media data, including user comments and the popularity of the cast and movie, employing sentiment analysis to predict IMDB scores [13]. They concluded that linear regression alone is insufficient for accurate box office prediction, as their manually created dataset lacked enough features to enhance precision. Some researchers have also explored third-party sources for data collection. Zhao Jie et al. found that using microblogs as a source for social media data mining, which integrates sentiment polarity and influence, improved their prediction model [14]. However, they did not fully account for real-time data changes and processing. Additionally, their small dataset limited the model's predictive possibilities. To overcome the limitations of unimodal embeddings, many scholars have synthesized multiple models. For instance, Singh Krishna Kumar et al. found that using multiple modalities, even on smaller datasets, could yield more successful predictions [15]. Nonetheless, expanding dataset scales—incorporating diverse age groups, cultures, and preferences—would likely improve results. Additionally, capturing real-time data is essential for continuously adapting the model to dynamic changes.

A better way to predict the box office is to combine machine learning methods with deep learning-based methods. For instance, Wang Dongqi et al. attempted to balance the two main methods: classical machine learning and deep learning[8]. By leveraging the different advantages and disadvantages of both traditional machine learning and human wisdom, they developed a novel model to address the prediction problem. However, determining success criteria remains challenging due to the enormous diversity[16]. Sahu Sandipan et al. proposed a new model that helps establish the success criteria for films. However, executing successful data mining and classifying film popularity into many classes posed challenges in their research. Thus, we can conclude that artificial intelligence can make promising progress by combining the two main methodologies in the future. Integrating both classical machine

DOI: 10.54254/2755-2721/116/20251729

learning-based methods and deep learning-based methods will contribute to developing better models and methodologies for box office prediction.

3. Discussion

3.1. General Processes of Box-office Prediction

Many box office predictions rely on social data and are made using artificial intelligence methods, following these general steps:

The process begins with collecting data from diverse sources. The available data is then processed based on specific criteria, followed by analysis and model development. Finally, multiple predictions are made using the established model.

3.1.1. Gathering The Data(GTA). GTA tends to collect diverse data from all the aspects of the film, range from official to third-party platforms to get the information sources. Then it needs to use website crawlers to extract all relevant data about the movies, time of the year, advertisement, existed languages and so on. In addition, it is essential to capture viewers' subjective emotions around the social media to convert it into sentiment data.

*3.1.2. Preprocessing The Data(PTA).*PTA tends to make full preparation of the data which could be access and extracted by the ML or DL process. On the one hand, this step could abandon numerous untuneful and irrelevant information. On the other hand, it is the high time to set and select relevant criteria for objective features and other process for the data.

*3.1.3. Developing Prediction Algorithm(DPA).*DPA tends to implement selected ML or DL algorithms for the next step EAE. Additionally, there are lots of programming languages benefit the algorithm implement, C++, java, python and so on. And there are a wide range of ML or DL tools that are useful too.

3.1.4. Experiments And Evaluation(EAE). EAE tends to use different prediction experiments range from different prediction algorithms, different parameters to different features. Different evaluations develop each experiments' advantages and disadvantages and the performance of the approach in relation to other approaches[17].

3.2. Comparison of Different Strengths between ML and DL

Discussion based on the table. The comparison of strengths and weaknesses between ML and DL is shown in Table 1. It demonstrates that their strengths and weaknesses vary from one aspect to another.

(1) Deep learning-based early box office prediction has a performance advantage when the dataset is significantly large; however, if the amount of data is smaller, a traditional machine learning-based approach is recommended.

(2) Using multiple sources of social media data for prediction will be an important method to enhance the precision and reliability of predictions. Different media publish data in various aspects and at different speeds, and based on the differences and commonalities of multiple sources, the reliability of the data can be further ascertained, and the adequacy of the data volume can be ensured. However, effectively handling social media data from multiple sources remains a significant challenge.

(3) Adopting multimodal data for prediction will be another development trend. Combining different modalities can improve prediction speed and accuracy, but managing multimodal data is still a substantial challenge.

(4) From the perspective of training cost and time investment, traditional machine learning is significantly better than deep learning. Traditional machine learning has a more established methodology, while deep learning applies to more demanding conditions.

DOI: 10.54254/2755-2721/116/20251729

(5) In terms of feature processing, deep learning can directly extract features from raw objective data, while traditional machine learning requires the use of processed subjective data. The nature of feature extraction thus determines the different prediction speeds and labor requirements.

(6) There are also distinctions between the two methods when they are further subdivided. There is little difference among traditional machine learning models, while there is a notable difference in the effectiveness of deep learning models with or without coding.

Table 1. Comparison of different strengths between ML and DL.

Differences	ML Advantages	DL Advantages
Time, economic costs	Less	More
source of information	Feature engineering	More complex dataset
Feature Processing	(1) ML need subjective feature (2) ML tend to transfer the feature into one-dimensional vectors	(1) DL need objective feature (2) DL tends to distinguish resemble feature (3) DL tend to process feature into two- dimensional image data time-effectively
Differences in respective internal methodologies	Different ML methodologies have no obvious difference with each other	Encoded one tend to perform better than non-encoded one

4. Conclusion

AI greatly benefits the issue of box office prediction. Early box office predictions based on social media are crucial for making intelligent decisions in movie investments and theater screening strategies. As a result, early box office predictions attract a wide range of researchers. In this paper, we reviewed the research on early box office prediction to gain insights and suggest directions for future research. First, the general process of social media-based early box office prediction is summarized. Secondly, two types of research methods are examined: traditional machine learning (ML) and deep learning (DL). After discussion, we conclude that deep learning-based methods offer greater flexibility. Furthermore, the study results indicate that deep learning is recommended when (1) the dataset is large, (2) the features are subjective, and (3) the embedding model has been coded. Machine learning is recommended when (1) budget and time are limited and (2) the features are objective. Additionally, more features such as cast, number of screens, and genre should be extracted from social media (Twitter, YouTube, Weibo, blogs, and the IMDb movie datasets) to improve predictive performance.

References

[1] M. Sudais, M. H. Khan, and A. J. Tabani, “Performance Prediction for IMDB Movies,” 2022, [Online]. Available: https://doi.org/10.21203/rs.3.rs-1243202/v1

[2] Y. Liao, Y. Peng, S. Shi, V. Shi, and X. Yu, “Early box office prediction in China’s film market based on a stacking fusion model,” Ann Oper Res, vol. 308, no. 1–2, pp. 321–338, 2022, doi: 10.1007/s10479-020-03804-4.

[3] Y. Ni, F. Dong, M. Zou, and W. Li, “Movie Box Office Prediction Based on Multi-Model Ensembles,” Information (Switzerland), vol. 13, no. 6, 2022, doi: 10.3390/info13060299.

[4] W. Lu, X. Zhang, and X. Zhan, “Movie Box Office Prediction Based on IFOA-GRNN,” Discrete Dyn Nat Soc, vol. 2022, pp. 1–4, 2022, doi: 10.1155/2022/3690077.

[5] T. Song, J. Huang, Y. Tan, and Y. Yu, “Using user- and marketer-generated content for box office revenue prediction: Differences between microblogging and third-party platforms,” Information Systems Research, vol. 30, no. 1, pp. 191–203, 2019, doi: 10.1287/isre.2018.0797.

[6] S. Sahu, R. Kumar, H. V. Long, and P. M. Shafi, Early-production stage prediction of movies success using K-fold hybrid deep ensemble learning model, vol. 82, no. 3. Multimedia Tools and Applications, 2023. doi: 10.1007/s11042-022-13448-0.

DOI: 10.54254/2755-2721/116/20251729

[7] R. A. Masrury, M. A. A. Saputra, A. Alamsyah, and M. A. S. Primantari, "A comparative study of Hollywood movie successfulness prediction model," 2019 7th International Conference on Information and Communication Technology, ICoICT 2019, pp. 1–5, 2019, doi: 10.1109/ICoICT.2019.8835385.

[8] A. Al Fahoum and T. A. Ghobon, "Performance Predictions of Sci-Fi Films via Machine Learning," Applied Sciences (Switzerland), vol. 13, no. 7, 2023, doi: 10.3390/app13074312.

[9] J. Park and C. Lim, "Predicting Movie Audience with Stacked Generalization by Combining Machine Learning Algorithms," Commun Stat Appl Methods, vol. 28, no. 3, pp. 217–232, 2021, doi: 10.29220/CSAM.2021.28.3.217.

[10] D. Wang et al., "A movie box office revenues prediction algorithm based on human-machine collaboration feature processing," Journal of Engineering Research (Kuwait), vol. 10, pp. 1–17, 2022, doi: 10.36909/jer.ICCSCT.19489.

[11] D. Menaga and A. Lakshminarayanan, "A Method for Predicting Movie Box-Office using Machine Learning," 2023 4th International Conference on Electronics and Sustainable Communication Systems, ICESC 2023 - Proceedings, pp. 1228–1232, 2023, doi: 10.1109/ICESC57686.2023.10192928.

[12] M. Y. Ko, J. L. Li, and C. C. Lee, "Learning minimal intra-genre multimodal embedding from trailer content and reactor expressions for box office prediction," Proc (IEEE Int Conf Multimed Expo), vol. 2019-July, pp. 1804–1809, 2019, doi: 10.1109/ICME.2019.00310.

[13] C. T. Madongo and T. Zhongjun, "A movie box office revenue prediction model based on deep multimodal features," Multimed Tools Appl, vol. 82, no. 21, pp. 31981–32009, 2023, doi: 10.1007/s11042-023-14456-4.

[14] I. Sindhu and F. Shamsi, "Prediction of IMDB Movie Score and Movie Success by Using the Facebook," 2023 International Multi-Disciplinary Conference in Emerging Research Trends, IMCERT 2023, vol. I, pp. 1–5, 2023, doi: 10.1109/IMCERT57083.2023.10075189.

[15] J. Zhao, F. Xiong, and P. Jin, "Enhancing Short-Term Sales Prediction with Microblogs: A Case Study of the Movie Box Office," Future Internet, vol. 14, no. 5, 2022, doi: 10.3390/fi14050141.

[16] K. K. Singh, J. Makhania, and M. Mahapatra, "Impact of ratings of content on OTT platforms and prediction of its success rate," Multimed Tools Appl, vol. 83, no. 2, pp. 4791–4808, 2024, doi: 10.1007/s11042-023-15887-9.

[17] I. S. Ahmad, A. A. Bakar, M. R. Yaakub, and S. H. Muhammad, "A Survey on Machine Learning Techniques in Movie Revenue Prediction," SN Comput Sci, vol. 1, no. 4, pp. 1–14, 2020, doi: 10.1007/s42979-020-00249-1.