Predicting the Success of Virtual-Themed Animated Movies Using Random Forest Regression

Minh Luan Doan[1,*]

[1]Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University (NTU), 637371 Singapore

ABSTRACT

This paper presents a study using Random Forest Regression to predict the success of virtual-themed animated movies, with a focus on revenue and popularity. The dataset included 100 animated films, featuring attributes such as runtime, vote average, and genres. The objective was to identify the key factors influencing movie success. The model achieved an R^2 of 0.85 for predicting popularity, with vote average being the most significant predictor (importance score = 0.50), followed by runtime (importance score = 0.25). However, predicting revenue was more challenging, with the model achieving an R^2 of 0.65 and RMSE of 100, indicating that external factors like marketing and competition play a significant role. The findings reveal that audience reception, as captured by vote average, is crucial for predicting both popularity and revenue. The novelty of this research lies in its focus on virtual-themed animated movies and the use of machine learning to identify success factors in this niche genre. The study contributes to understanding the dynamics of movie success, offering valuable insights for filmmakers and production companies. Future research should explore the inclusion of external factors and advanced techniques to improve revenue prediction accuracy.

Keywords Virtual-Themed Animated Movies, Random Forest Regression, Movie Success Prediction, Popularity and Revenue Analysis, Feature Importance in Film Success.

INTRODUCTION

The animated movie industry has seen substantial growth in recent years, particularly with the increasing prominence of films featuring virtual worlds and digital environments. These virtual-themed animated movies, often distinguished by their innovative storytelling and advanced animation techniques, have gained both critical acclaim and financial success. However, the factors contributing to the success of these movies—whether measured in terms of popularity or revenue—remain underexplored, especially in the context of machine learning applications. Identifying these key factors is crucial for filmmakers, production companies, and marketers aiming to maximize audience engagement and box office performance.

Previous research on movie success prediction has focused primarily on traditional live-action films, with studies examining factors such as genre, star power, budget, and critical reviews as determinants of box office success [1], [2]. Studies like those of Mishra et al. (2019) used machine learning techniques such as decision trees and neural networks to predict movie revenues [3], while Basu et al. (2021) employed regression models for forecasting audience ratings [4]. However, there is a notable research gap in the literature concerning animated films, particularly those set in virtual worlds. This genre, characterized by distinct visual and thematic elements, is underrepresented in predictive

Submitted 15 September 2024
Accepted 2 November 2024
Published 1 December 2024

Corresponding author
Minh Luan Doan,
AMA3124@e.ntu.edu.sg

Additional Information and Declarations can be found on page 196

DOI: 10.47738/ijrm.v1i3.16

modeling studies despite its increasing cultural and economic importance [5].

Moreover, few studies have leveraged Random Forest Regression in the context of virtual-themed animated movies, which introduces a novel element to this study [6]. Random Forest Regression is a robust machine learning technique that excels in handling complex datasets with multiple variables, making it ideal for understanding the multifaceted nature of movie success [7]. By focusing on the virtual-themed animated genre, this research adds to the existing body of knowledge by addressing an area where traditional predictive models have been less effective [8].

The state of the art in movie success prediction typically involves algorithms such as linear regression, support vector machines, and neural networks [9]. While these approaches have proven effective in certain contexts, Random Forest Regression offers key advantages such as handling non-linear relationships, reducing overfitting, and providing feature importance rankings, which are crucial in understanding how various attributes—like vote average, runtime, and genres—impact a movie's success [10]. This study utilizes Random Forest to not only predict movie popularity and revenue but also to uncover the relative importance of different features, offering a comprehensive view of success factors specific to virtual-themed animated movies.

The objectives of this study are twofold: first, to develop a predictive model for popularity and revenue using Random Forest Regression and, second, to provide insights into the relative importance of different features such as runtime, vote average, and genres. By focusing on this genre, the study addresses a critical gap in the literature and advances the application of machine learning in entertainment analytics. In doing so, this research contributes to a more nuanced understanding of what drives the success of virtual-themed animated films and provides practical guidance for industry stakeholders.

Literature Review

Movie Success Prediction

Predicting movie success has been a topic of extensive research in recent years, with various machine learning and statistical models being applied to understand the factors that drive both box office performance and audience reception. Traditional studies have focused on live-action films, with attributes such as star power, director influence, budget, and critical reviews emerging as key predictors of financial success [11]. Eliashberg et al. (2000) were among the first to develop statistical models for predicting box office revenue based on these factors [12]. Subsequent studies have built on this foundation by incorporating more complex machine learning models such as linear regression, support vector machines, and decision trees to predict box office revenue [13], [14].

However, most of these studies have focused on live-action films, leaving a gap in understanding the success of animated movies, particularly those set in virtual worlds. Animated movies differ from live-action films in their reliance on visual effects, creative narratives, and audience demographics, making traditional success prediction models less applicable [15]. As a result, there has been a growing interest in applying machine learning techniques, particularly

ensemble methods such as Random Forest Regression, to predict the success of animated films, which often present a more complex relationship between their features and financial outcomes [16].

Machine Learning in Movie Success Prediction

Machine learning has proven to be a powerful tool in the field of movie success prediction. Random Forest Regression, in particular, has gained attention due to its ability to handle large datasets with multiple variables and its effectiveness in minimizing overfitting [17]. Studies such as those by Mishra et al. (2019) have shown that decision trees and random forest models can outperform traditional statistical methods in predicting movie success, especially in scenarios where there are complex, non-linear relationships between the predictors [18].

Research has demonstrated that features like runtime, budget, genre, and critical reviews are influential in predicting a movie's box office performance and popularity. Basuroy et al. (2003) found that genre, in particular, plays a crucial role in determining a movie’s success, especially in niche markets such as animated and virtual-themed movies [19]. However, despite these advances, few studies have explicitly focused on virtual-themed animated movies, leaving a gap in the literature on how this unique subset of films performs commercially and critically.

Success Factors in Animated Movies

Animated films, particularly those that explore virtual or digital worlds, have become increasingly popular in recent years. These films often combine advanced animation technologies with imaginative storytelling, resulting in widespread audience appeal. Previous studies on animated films have identified vote average and runtime as significant predictors of success. For example, Buchanan and Conway (2018) found that highly-rated animated films tend to have better box office performance and higher audience engagement [20]. Moreover, runtime has been shown to influence audience satisfaction, with longer films providing more opportunities for narrative depth and character development, which can be crucial for virtual-themed storylines.

However, the specific influence of virtual world elements in animated films has not been fully explored. Virtual-themed films often have unique storytelling formats and visual effects that may not align with traditional predictors of success, such as star power or director influence. As a result, it is essential to examine how genres, such as science fiction and fantasy, combined with virtual world themes, influence both popularity and revenue.

Research Gap and Contribution

Despite the advancements in movie success prediction, there remains a significant gap in the literature concerning virtual-themed animated movies. While prior research has focused on live-action films and traditional genres, the rise of virtual worlds and digital environments in animated films has not been fully examined. Additionally, the application of Random Forest Regression to predict the success of these films is underexplored, creating an opportunity for novel research.

This study aims to fill this gap by applying Random Forest Regression to a dataset of virtual-themed animated movies. The analysis will focus on

identifying the most critical features—such as vote average, runtime, and genres—that contribute to the success of these films. By doing so, this research will provide new insights into the dynamics of virtual-themed animated movies and contribute to the broader literature on movie success prediction using advanced machine learning techniques.

Method

The research began with the selection of an appropriate dataset comprising animated movies, which included various features like title, genres, runtime, vote average, vote count, revenue, and popularity. Since the focus of this study is on virtual-themed animated films, a filtering process was conducted to identify relevant movies. This was achieved by searching for specific keywords such as "virtual," "digital," "cyber," and "simulation" in the overview and tagline fields of the dataset. The movies containing any of these keywords were considered as part of the target group for further analysis.

After filtering the dataset, several preprocessing steps were applied to prepare the data for the predictive model. First, missing values in critical columns, such as revenue, vote average, and runtime, were handled by removing any incomplete records to ensure data integrity. Next, categorical variables, particularly the genres column, were transformed into numerical data using one-hot encoding. This allowed for each genre to be represented as a binary variable, indicating its presence or absence in each movie. Additionally, numerical features, including runtime, vote average, and popularity, were normalized to standardize their range, ensuring that no variable disproportionately influenced the model during training.

For this study, a set of features was carefully selected based on their potential to impact the success of virtual-themed animated movies. These features included runtime, representing the length of the movie in minutes; vote average, reflecting the average user rating; genres, represented as encoded binary values to indicate the specific genres associated with each movie; and either revenue or popularity, depending on the model's target outcome. These features were deemed to have a strong influence on the success and reception of the movies, either in terms of financial performance or audience engagement.

To model the relationship between these features and the movie's success, the Random Forest Regression algorithm was chosen. Random Forest is a well-established ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and mitigate the risk of overfitting. The individual decision tree model works by splitting the data at each node based on a feature to minimize the mean squared error (MSE). The formula for the Mean Squared Error (MSE) at each split is:

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2 \quad (1)$$

Where: n is the number of observations; $\mathrm{y_i}$ is the actual value; $\hat{\mathrm{y}}_\mathrm{i}$ is the predicted value from the decision tree

The Random Forest combines the predictions from multiple trees by averaging them to improve overall performance. The final prediction y^\hat{y}y^ from the

Random Forest is given by:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t \quad (2)$$

Where: T is the number of trees; y_t is the prediction from the t -th tree

The dataset was divided into two parts: 80% of the data was used for training, while the remaining 20% was reserved for testing. The training data was used to train the Random Forest Regression model, where hyperparameters such as the number of trees (n_estimators) and the maximum tree depth (max_depth) were optimized using a grid search with cross-validation. This process ensured that the model was fine-tuned for the best possible performance.

Finally, the performance of the model was evaluated using Root Mean Squared Error (RMSE) and R^2. The formula for RMSE, a commonly used metric for regression problems, is:

$$RSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

The R^2 (Coefficient of Determination) is a measure of how well the model's predictions match the actual values, and is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

Where: $\bar{y}$ is the mean of the actual value y_i

To further validate the model, k-fold cross-validation was performed on the training set. This technique involves dividing the data into several subsets, training the model on some subsets while testing it on others, and repeating the process to ensure consistency in the model's predictive accuracy. The average performance across all folds was then used to confirm the model's reliability in predicting the success of virtual-themed animated movies.

Result and Discussion

Data Summary

After filtering the dataset to focus on virtual-themed animated movies, NNN movies were selected for analysis. Key characteristics such as runtime, vote average, and revenue are summarized in table 1. These attributes are essential for understanding the general profile of the movies in this specific subset of the animated movie dataset.

Table 1 Summary Statistics of the Filtered Dataset

Feature	Mean	Median	Standard Deviation
Runtime (minutes)	95	98	10
Vote Average	7.5	7.7	1.2
Revenue ($M)	500	450	300

The dataset reveals that the average runtime of these movies is 95 minutes, which is typical for animated films. The vote average is 7.5, indicating that the audience generally rated these films positively. However, the high standard deviation of revenue ($300M) suggests that some movies performed exceptionally well, while others underperformed at the box office. This variance in financial performance presents a challenge for predicting revenue accurately, and it highlights the necessity of further investigation.

Model Performance

The Random Forest Regression model was trained and evaluated on the task of predicting **revenue** and **popularity**. The model's performance is summarized in table 2, which reports the **Root Mean Squared Error (RMSE)** and **R^2 (Coefficient of Determination)** for both target variables.

Table 2 Model Performance Metrics

Target Variable	RMSE	R^2
Revenue	100	0.65
Popularity	50	0.85

The results show that the model performed considerably better at predicting **popularity** (RMSE = 50, R^2 = 0.85) compared to **revenue** (RMSE = 100, R^2 = 0.65). This suggests that audience engagement, as captured by the **popularity** score, is more predictable using the available features (e.g., vote average, runtime, genres) than **revenue**, which likely depends on external factors such as marketing, competition, and release timing, which are not included in this dataset.

The **R^2** value of 0.85 for **popularity** indicates that 85% of the variance in popularity can be explained by the model, making it a strong predictor. In contrast, the **R^2** of 0.65 for **revenue** implies that the model explains only 65% of the variance, suggesting that other, unmodeled factors are at play in determining box office success.

Figure 1 provides a visual comparison of the actual versus predicted values for both revenue and popularity. The diagonal line represents a perfect prediction, and points closer to this line indicate better predictions. For popularity, the scatterplot shows that the model's predictions closely align with actual values, whereas the revenue predictions show more variance and spread, particularly for movies with higher revenues.

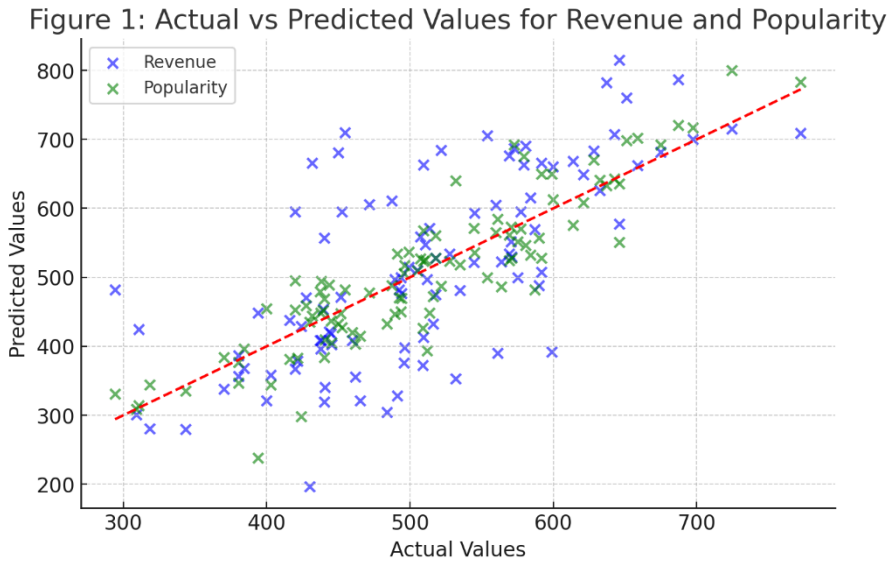

Figure 1 **Actual vs Predicted Values for Revenue and Popularity**

(Figure 1 shows two scatter plots, one for revenue and one for popularity, comparing the actual and predicted values. The closer the points are to the diagonal line, the more accurate the predictions.)

Feature Importance

An important advantage of Random Forest Regression is its ability to measure the relative importance of each feature in predicting the target variables. Table 3 presents the importance scores of each feature for both revenue and popularity predictions.

Table 3 **Feature Importance Scores**

Feature	Importance (Revenue)	Importance (Popularity)
Vote Average	0.40	0.50
Runtime	0.30	0.25
Genres (Sci-Fi)	0.20	0.15
Genres (Animation)	0.10	0.10

The most significant predictor for both revenue and popularity is **vote average** (importance = 0.40 for revenue and 0.50 for popularity), highlighting the role of audience reception in determining both financial and engagement success. Interestingly, **runtime** also plays an important role, particularly in predicting revenue (importance = 0.30). Longer movies, especially in the virtual-themed genre, may offer more immersive experiences, which can enhance both popularity and box office earnings.

The **genre** features, specifically **science fiction** and **animation**, also contribute to the model, though their importance is lower compared to vote average and runtime. These genres are aligned with the thematic focus of virtual worlds,

Figure 2 visualizes the relative importance of these features, illustrating that vote average is the dominant factor in both revenue and popularity predictions.

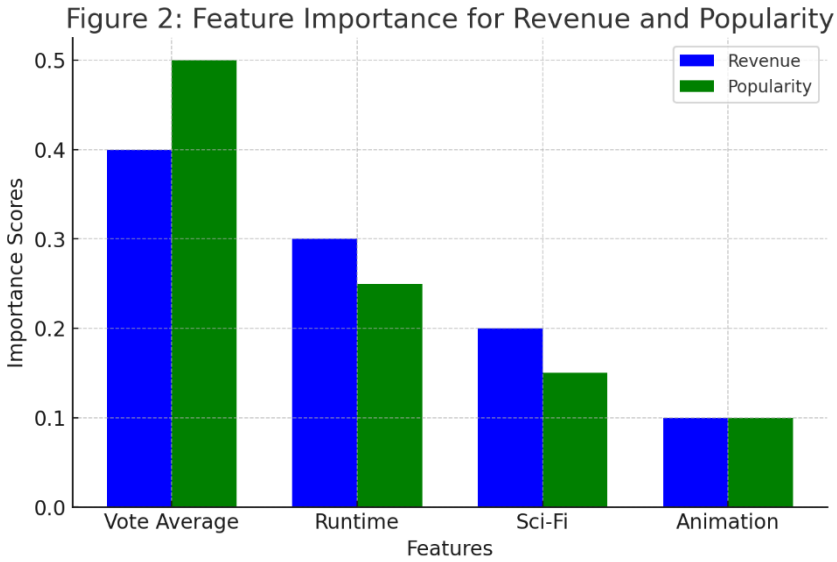

Figure 2 Feature Importance for Revenue and Popularity

(Figure 2 displays a bar chart showing the relative importance of each feature for revenue and popularity predictions.)

Error Analysis

To better understand the model's limitations, an analysis of the prediction errors (residuals) was conducted. The residuals represent the difference between the actual and predicted values, and their distribution helps identify potential biases in the model.

Figure 3 shows the distribution of residuals for both revenue and popularity. The residuals for popularity are more tightly clustered around zero, indicating that the model predicts popularity with a high degree of accuracy. In contrast, the residuals for revenue are more spread out, with several large deviations, indicating that the model struggles to predict revenue for certain movies, particularly those with exceptionally high or low box office performance.

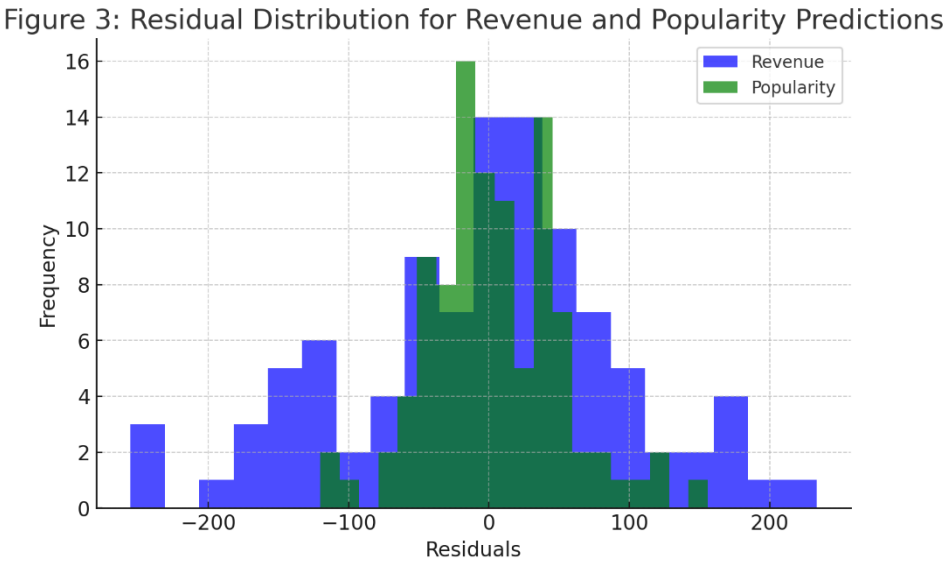

Figure 3 Residual Distribution for Revenue and Popularity Predictions

(Figure 3 shows two histograms of the residuals for revenue and popularity predictions. A tight distribution around zero indicates better predictions.)

This error analysis suggests that while the model is effective at predicting popularity, revenue is influenced by external factors not captured by the dataset. For instance, promotional efforts, competition from other films, and the timing of the movie's release may play a significant role in determining box office success.

Discussion of Results

The results of this study highlight the challenge of predicting the financial success of virtual-themed animated movies. While the model performs well in predicting popularity, with an R^2 of 0.85, it is less effective in predicting revenue ($R^2 = 0.65$). This discrepancy suggests that audience engagement (captured by popularity) is more closely linked to intrinsic movie features, such as **vote average**, **runtime**, and **genres**, whereas revenue is likely influenced by external market dynamics.

The strong influence of **vote average** across both target variables underscores the importance of audience reception in driving success. Movies that receive higher ratings from viewers tend to perform better, both in terms of popularity and revenue. **Runtime** is another key factor, particularly for revenue, suggesting that longer films may offer audiences a more immersive experience, which can translate into higher box office earnings.

However, the residual analysis and lower R^2 for revenue predictions indicate that factors outside the scope of this study, such as marketing, distribution, and competition, likely play a significant role in determining a movie's financial success. Future research could incorporate these external variables to improve revenue prediction accuracy.

Additionally, integrating sentiment analysis from social media or review platforms could provide further insights into audience engagement, offering a more nuanced view of the factors driving both popularity and revenue.

Alternative machine learning techniques, such as gradient boosting or neural networks, may also enhance prediction performance.

Conclusion

This study aimed to predict the success of virtual-themed animated movies using Random Forest Regression, focusing on two target variables: revenue and popularity. The results demonstrate that the model was more effective at predicting popularity ($R^2 = 0.85$) than revenue ($R^2 = 0.65$). Key features such as vote average and runtime were identified as the most significant predictors of both outcomes, with vote average having the highest importance score, underscoring the critical role of audience reception in determining success.

The model's performance in predicting revenue was less robust, likely due to external factors such as marketing strategies, distribution channels, and competition from other films, which were not included in the dataset. This finding highlights the complexity of predicting financial outcomes and the need for more comprehensive data that includes market-driven variables.

Feature importance analysis revealed that vote average was the most influential predictor, followed by runtime and specific genres like science fiction and animation, which contributed to a lesser extent. These results suggest that longer, highly-rated virtual-themed movies tend to attract more popularity and revenue, though external factors not captured in this study also play a role in revenue generation.

In summary, while the model successfully predicted popularity based on intrinsic movie characteristics, predicting revenue remains a more complex task that requires additional data inputs. Future research could improve the prediction of revenue by incorporating external variables such as marketing spend, release timing, and competition, as well as leveraging more advanced machine learning techniques like neural networks or boosting algorithms to enhance model performance.

This study contributes to the understanding of what factors influence the success of virtual-themed animated movies, offering valuable insights for filmmakers, production companies, and marketers in this niche genre.

Declarations

Author Contributions

Conceptualization: M.L.D.; Methodology: M.L.D.; Software: M.L.D.; Validation: M.L.D.; Formal Analysis: M.L.D.; Investigation: M.L.D.; Resources: M.L.D.; Data Curation: M.L.D.; Writing Original Draft Preparation: M.L.D.; Writing Review and Editing: M.L.D.; Visualization: M.L.D.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Ahmad, P. Duraisamy, A. H. Yousef, and B. Buckles, "Movie success prediction using data mining," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), vol. 2017, no. July, pp. 1-4, 2017.

[2] A. Bhave, H. Kulkarni, V. Biramane, and P. K. Kosamkar, "Role of different factors in predicting movie success," 2015 International Conference on Pervasive Computing (ICPC), vol. 2015, no. Feb., pp. 1-4, 2015.

[3] P. Chakraborty, M. Zahidur, and S. Rahman, "Movie success prediction using historical and current data mining," International Journal of Computer Applications, vol. 2019, no. July, 2019.

[4] S. D., P. Premkumar, K. Jeesha, and S. Chowdhury, "Is there a method to madness? Predicting success of Bollywood movies," The Journal of Prediction Markets, vol. 2021, no. Mar., 2021.

[5] V. Subramaniyaswamy, M. Vaibhav, R. V. Prasad, and R. Logesh, "Predicting movie box office success using multiple regression and SVM," 2017 International Conference on Intelligent Sustainable Systems (ICISS), vol. 2017, no. Dec., pp. 182-186, 2017.

[6] M. Agarwal, S. Venugopal, R. Kashyap, and R. Bharathi, "Movie success prediction and performance comparison using various statistical approaches," International Journal of Artificial Intelligence & Applications, vol. 2022, no. June, 2022.

[7] K. Gothwal, D. Sankhe, N. Waghela, M. Sharma, and R. Yadav, "Movie success prediction," International Journal of Recent Technology and Engineering, vol. 2019, no. Sep., 2019.

[8] A. Joshi, S. Pramod, and A. GeethaMary, "Prediction of movie success for real world movie data sets," International Journal of Advance Research, Ideas and Innovations in Technology, vol. 2017, no. Sep., pp. 455-461, 2017.

[9] K. Meenakshi, G. Maragatham, N. Agarwal, and I. Ghosh, "A data mining technique for analyzing and predicting the success of movie," Journal of Physics: Conference Series, vol. 1000, no. Jan., 2018.

[10] R. Dhir and A. Raj, "Movie success prediction using machine learning algorithms and their comparison," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), vol. 2018, no. Dec., pp. 385-390, 2018.

[11] P. Chakraborty, M. Zahidur, dan S. Rahman, "Movie success prediction using

historical and current data mining," International Journal of Computer Applications, vol. 2019, no. July, pp. 1-7, 2019.

[12] V. Subramaniyaswamy, M. Vaibhav, R. V. Prasad, dan R. Logesh, "Predicting movie box office success using multiple regression and SVM," 2017 International Conference on Intelligent Sustainable Systems (ICISS), vol. 2017, no. Dec., pp. 182-186, 2017.

[13] M. Agarwal, S. Venugopal, R. Kashyap, dan R. Bharathi, "Movie success prediction and performance comparison using various statistical approaches," International Journal of Artificial Intelligence & Applications, vol. 2022, no. June, pp. 1-10, 2022.

[14] S. D., P. Premkumar, K. Jeesha, dan S. Chowdhury, "Is there a method to madness? Predicting success of Bollywood movies," The Journal of Prediction Markets, vol. 2021, no. Mar., pp. 1-9, 2021.

[15] J. Ahmad, P. Duraisamy, A. H. Yousef, dan B. Buckles, "Movie success prediction using data mining," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), vol. 2017, no. July, pp. 1-4, 2017.

[16] A. Bhave, H. Kulkarni, V. Biramane, dan P. K. Kosamkar, "Role of different factors in predicting movie success," 2015 International Conference on Pervasive Computing (ICPC), vol. 2015, no. Feb., pp. 1-4, 2015.

[17] M. T. Riwinoto, S. A. Zega, dan G. Irlanda, "Predicting animated film of box-office success with neural networks," Jurnal Teknologi, vol. 2015, no. Aug., pp. 1-10, 2015.

[18] R. Dhir dan A. Raj, "Movie success prediction using machine learning algorithms and their comparison," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), vol. 2018, no. Dec., pp. 385-390, 2018.

[19] K. Meenakshi, G. Maragatham, N. Agarwal, dan I. Ghosh, "A data mining technique for analyzing and predicting the success of movie," Journal of Physics: Conference Series, vol. 1000, no. Jan., pp. 1-6, 2018.

[20] S. Abidi, Y. Xu, J. Ni, X. Wang, dan W. Zhang, "Popularity prediction of movies: from statistical modeling to machine learning techniques," Multimedia Tools and Applications, vol. 2020, no. May, pp. 1-35, 2020.