Machine Learning Methods for Predicting the Popularity of Movies

David Opeoluwa Oyewola[1], Emmanuel Gbenga Dada[2,*]

[1] Department of Mathematics and Statistics, Federal University Kashere, Gombe, Nigeria
Email: davidoyewole@fukashere.edu.ng
[2] Department of Mathematical Sciences, University of Maiduguri, Maiduguri, Nigeria
Email: gbengadada@unimaid.edu.ng
*Corresponding Author: Emmanuel Gbenga Dada, Email: gbengadada@unimaid.edu.ng

How to cite this paper: David Opeoluwa Oyewola, Emmanuel Gbenga Dada (2022). Machine Learning Methods for Predicting the Popularity of Movies. Journal of Artificial Intelligence and Systems, 4, 65–82. https://doi.org/10.33969/AIS.2022040105.

Received: September 20, 2022
Accepted: October 13, 2022
Published: October 17, 2022

Abstract

The movie industry has grown into a several billion-dollar enterprise, and there is now a ton of information online about it. Numerous machine learning techniques have been created by academics and can produce effective classification models. In this study, different machine learning classification techniques are applied to our own movie dataset for multiclass classification. This paper's main objective is to compare the effectiveness of various machine learning techniques. This study examined five methods: Multinomial Logistic Regression (MLR), Support Vector Machine (SVM), Bagging (BAG), Naive Bayes (NBS) and K-Nearest Neighbor (KNN), while noise was removed using All K-Edited Nearest Neighbors (AENN). These techniques all utilize previous IMDb dataset to predict a movie's net profit value. The algorithms predict the profit at the box office for each of these five techniques. Based on the dataset used in this paper, which consists of 5043 rows and 14 columns of movies, this study evaluates the performance of all seven machine learning techniques. Bagging outperformed other machine learning techniques with a 99.56% accuracy rate.

Keywords

Multinomial Logistic Regression (MLR), Support Vector Machine (SVM), Bagging (BAG), Naive Bayes (NBS), Movie popularity

1. Introduction

The World Wide Web is currently the most reliable resource for information, and it's amazing how much data every profession is releasing online and how much faster and more effective it is. The film industry also generates a large amount of data online on actors, directors, studios, reviewers' scores, ratings, and many more, making it easier for scientists to analyze the data, track, and discover the unseen patterns within this vast data about films [1]. It takes a lot of work for filmmakers to produce a popular film. Due to the diverse tastes of the audiences, focusing solely on a few aspects of a film, such as its genre, casting, and stars, may not be sufficient. A film's popularity is influenced by both conventional and unconventional elements, including the director, well-known actors/actresses, genre, and budget [2]. Non-conventional factors include number of people that watched the movie clips on YouTube, likes on social media, and number of fans following the movie. The term "success" for a movie is a subjective one; sometimes films are deemed successful based on their global box office haul, while others—while underperforming commercially—can still be deemed successful due to their favorable reviews from critics, high ratings, and popularity [3].

Movies perform an important role as a driver of the process of learning to behave in a way that is acceptable to society. Movies are known to habitually work without restrictions in opposition to the virtues and morals of traditional communal fundamentals [4]. Films are identified as one of the essential bedrock of society nowadays which influence the advancement of any society. Since its introduction about a century ago, movies have gained recognition as a source of amusement. This trend has been extensively observed in the sales of film products, movie licenses, the fan base of millions of people worldwide, hero-worshiping, and romanticizing film superstars. Films are considered to be an agent of commentary on social issues, assessment, and condemnation.

The Motion Picture Association of America (MPAA) asserts that the expansion and advancement of the movie business are a worldwide experience. Because of the impact the film industry has on the country’s economic development, several research have been undertaken by academics using machine learning approach to predicted the success of movies. These studies have a significant impact on the film industry due to the predictive viewpoint [5]. Every year, many films are released by the film industry. According to the research, the movie sector in the United States generates profits of up to $25.9 billion in the year 2020, and nearly every movie costs about $100 million to produce. Despite these costs and the uncertainty surrounding whether a film will be successful or not, however, are there still some confusions and uncertainties [6].

The movie business is among the industries that generates the most money from a business standpoint. Furthermore, a single film's success can bring in huge profit for a production company, and filmmakers are eagerly eager to make a profit from the films through timely predictions that, as a film gains popularity with the wider populace, it will also generate earnings from the immediate society [7]. Many individuals enjoy watching movies as a hobby and are passionate about it. People appreciate watching the movie in theaters, and it is a huge form of recreation. The film business makes hundreds of films in a variety of genres (such as action, adventure, documentary, drama, animation, comedy, mystery, fantasy, crime, biography, sci-fi, horror, romance, thriller, western, game show, family, music, film-noir and history) every year owing to the preferences of the diverse customer base [8].

Hollywood is the home of instinct, as seen by the vast majority of diversely interesting and topical films that are produced annually in the U. S. The producer is still unsure if a film will be successful or not, which gives rise to the idea of making success predictions for films before they are released [6, 9]. According to Cizmeci and Oguducu, a film's performance at the ticket sales may be aided by disclosing the important details prior to release. To make a film a success, the filmmaker and other members of the production team could make wise choices. For instance, if a film is a success, more people may view it in other cinemas, which would undoubtedly raise revenue [10].

Many research have been done in the effort to predict the popularity of films. Several of these studies incorporate user ratings majority of the time, as well as any predictions made by users on Instagram, Facebook pages, Twitter and YouTube. Conversely, there has only been a small amount of research done on predicting movies using factors such release dates, Oscar-winning actors, producers, studios, and duration [8]. Tang and colleagues evaluated IMDb and DouBan listings for films in the same genre as part of their study. Due to the limited amount of data, the preliminary results were unable to provide strong proof for the impact of the language other than English on the popularity of the movie. Afterward, they found the positive and negative sentiments, which can be taken as a robust indication of the recommendation and could help in predicting the popularity of a movie [11].

Movie genres, according to Wang and Zhang, play a significant part in a film's attractiveness since the movie business makes selections about the kinds of movies that viewers from various ethnic groups enjoyed, rated, and preferred. The ultimate purpose of the film industry is to create movies that produce income, and this depends on different market sectors and consumer preferences [12]. Netflix's machine learning based recommender system is the perfect example of the dominance of big data analytics/mining in the movie industry as far as the prediction of the most anticipated and probable movies is involved, as the system perfectly predicts which specific movie a particular consumer desires to watch afterwards [13].

In contrast, the abundance of data about the films on the internet encourages researchers to investigate knowledge discovery in data mining, machine learning and deep learning. The film industry and filmmakers are skeptical of the likelihood that the movie will become popular and profitable in the coming years. They constantly consider how to promote the film, on whom to focus, when to release the film, and how to promote it. It is for this purpose that a film's pre-release prediction is of highest importance to the filmmakers [13, 6].

While they made significant contributions to the advancement of prediction accuracy, the main goal of previous studies was to present new machine learning algorithms and assess their effectiveness alone. To increase and boost prediction accuracy, numerous elements and perspectives could be taken into further consideration. For instance, it is important to discover the unexpected, strange, and obscure features. The accuracy and scalability of machine learning algorithms can also be improved by feature extraction from core ones and feature selection, which are two more ways [5].

This study, which was inspired by these earlier investigations, tries to use and extract the pertinent information from the IMDb data in order to better comprehend the appeal of a particular film. In this work, we concentrate on the feature component strategy for improving prediction accuracy. In order to evaluate which model is best at resolving the regression issue, we also introduced confidence interval of 95% to ascertain the effectiveness of the machine learning models. The models can find patterns in the data that represent components that can be used to predict the future. It can also determine which predictors are reliable for predicting how well a movie will do. The enormous movie data can also be used to collect more features by adjusting the input settings and criteria. The major contributions of this work include:

I. A survey of machine learning algorithms that can be used for the prediction of movies success rate was presented.

II. Ascertained the most appropriate predictive model for automatic movie ratings on the IMDb dataset and predict the profit at the box office.

III. Developed a machine learning models for the prediction of movies popularity using Multinomial Logistic Regression (MLR), Support Vector Machine (SVM), Bagging (BAG), Naive Bayes (NBS), and K-Nearest Neighbor (KNN) machine learning models. While noise was removed using All K-Edited Nearest Neighbors (AENN).

IV. Evaluation of the ability of the proposed models for movies popularity prediction was done using different metrics such as accuracy (A_c), sensitivity (S_e), specificity (S_p), positive predicted value (PPV), negative predicted value (NPV), area under the curve (AUC), confidence interval (CI), Kappa (K) and 95% confidence interval.

This is how the remainder of the paper is organized. The analysis of prior research on movie success prediction is covered in Section 2. The methods used in this study are detailed in Section 3. Section 4 discusses several performance measures and examines the predictive performance of the created prediction model. In Section 5, we conclude by giving the reader some final views and suggestions for future study.

2. Related Works

Several work has been done by researchers who have used data from different sources to predict the success of movies. Jaiswal and Sharma [14] applied Bagging machine-learning model to build a model have the capacity to predict the success or failure of a Bollywood film, prior to its release. The researchers made use of data collected from different sources such as Cinemalytics, Box Office India, YouTube and Wogma. The downside of their approach is that the performance of the proposed Bagging model is low. Quader et al. [15] used a decision support system to predict the success of film investment sector. The proposed system uses historical data from different sources like IMDb, Rotten Tomato, Box Office Mojo and Meta Critic to predict the degree of success of a movie. The shortcoming of the proposed model is that the performance is poor as both post and pre-release features achieved low prediction accuracy. Moreover, there is need to use more state-of-the-art metric to validate the performance of the model as confusion matrix and accuracy used to assess the effectiveness of the model is not enough.

Lee et al. [16] investigated the effectiveness of several machine-learning techniques for movie prediction. The advantage of the proposed model Cinema Ensemble Model (CEM) is that it produced a better prediction result compared to other machine learning algorithms that was investigated. However, the performances of the proposed model is still fairly low. Marović et al. [17] did a study on machine-learning techniques for automatic movie ratings. The authors made use of IMDb film database which is accessible to the public. The method produced results that is better than many selected baseline techniques. However, the downside of their work is the poor performance of the model. Moreover, the performance metric used is not enough to validate the effectiveness of the proposed model.

Jernbäcker and Pojan [18] investigated the possibility of classifying film rating and box office revenue using metadata that are accessible prior to release. Metadata gotten from the internet was used for creating a model to predict movie rating. The shortcoming of their approach is that the model performed poorly. Moreover, the authors did not use any metric apart from ROC to evaluate the performance of the proposed model. Bristi, Zaman, and Sultana [19] proposed machine-learning model for predicting the success rate of dataset created from Hollywood film list from

Wikipedia and their score from IMDb film rating website. The model gives good classification measures with the data set. The performance of the system is good but there is need to validate the claims of the author by using more performance metrics. Meenakshi et al. [20] applied data mining approach predict the success rate of movies. Their technique focused on selection of important feature to predict the success of movies. The work indicated that having big budget does not imply that a movie will get high rating. The downside of the proposed techniques is their relatively low performance of k-means, RPART and Decision Tree used in the work.

Lee, Kim and Cheong [21] applied contextual word embedding models like BERT and ELMo, to predict the success of a story with reference to their value and attractiveness. The author opted for their approach because it does not need large dataset to train and test the model. The main disadvantage of the proposed approach is the relatively poor performance of the model in terms of accuracy, recall, precision and F1 score. Abidi et al. [22] investigated the effectiveness of several machine-learning algorithms for predicting the success of a movie. There was less emphasis on movie's data and features. Among the five machine-learning algorithm investigated, GLM have the best performance in term of prediction accuracy and RMSE compared to other algorithms considered in the study. The shortcoming of the proposed technique is that the performance is relatively low.

3. Methodology

3.1. Movie Dataset

The dataset consists of 5043 rows and 14 columns. The columns include: director_name, duration actor_1_name, actor_2_name, actor_3_name, genre, movie_title, num_voted_users, movie_imdb_link, num_user_for_reviews, language, country, title_year and imdb_score. The genre columns were the targets for this study. The numeric column was later extracted from the original movie data due to the nature of machine learning. The extracted numeric values features of the movie data are duration, num_voted_users, num_user_for_reviews, title_year, imdb_scores while the genres which is the targets for this study consists of action=0, adventure=1, documentary=2, drama=3, animation=4, comedy=5, mystery=6, fantasy=7, crime=8, biography=9, sci-fi=10, horror=11, romance=12, thriller=13, western=14, game show=15, family=16, music=17, film noir=18, history=19. Table 1 is the movie dataset used in this research.

Table 1 Movie dataset

Terms	Meaning
Genres	action=0, adventure=1, documentary=2, drama=3, animation=4, comedy=5,mystery=6, fantasy=7, crime=8, biography=9, sci-fi=10, horror=11, romance=12, thriller=13, western=14, game show=15, family=16, music=17, film noir=18, history=19
Duration	Duration of the movies
num_voted_users	Number users voted
num_user_for_reviews	Number users for reviews
title_year	Title year
imdb_score	Internet movie database score

3.2. Machine Learning

Machine learning is utilizing in this research to predict genres in movie dataset. The target genres values consist of action, adventure, documentary, drama, animation, comedy, mystery, fantasy, crime, biography, sci-fi, horror, romance, thriller, western, game show, family, music, film-noir and history. This approach will enable the classification of 20 genres for movie lovers. Five machine learning algorithms were utilized in this research while All K-Edited Nearest Neighbors (AENN) were used to filter the noise from movie dataset.

3.2.1. Multinomial Logistic Regression (MLR)

Logistic regression analysis is used in case of two-category dependent variable while multinomial logistic regression analysis is used to explain the cause-and-effect relationship between the independent variables and the dependent variable in case of dependent variable has at least three and more categories. Multinomial logistic regression model is an expanded version of the two-category model such as binary model [23]. Multinomial logistic regression can be defined as shown in Equation (1):

$$\pi(x) = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}} \qquad (1)$$

Where $\pi(x)$ defines the natural logarithm of the odds ratio, α and β signify the coefficients of parameters and x represents the independent variables.

3.2.2. Support Vector Machine (SVM)

SVM method classifies both linear and nonlinear data. A nonlinear mapping is utilized by SVM for converting the primary training set into an upper-level size. SVM examines for the linear optimal separating hyperplane in this new size like a decision border by which the tuples of one class from another are being split.

The data from two classes can be separated by a hyperplane which uses a proper nonlinear mapping to an upper dimension. This hyperplane is used to form support vectors that are important training vectors and margins. Contrary to the other methods, they are highly robust for overfitting [24].

3.2.3. Bagging (BAG)

Decisions taken from different learners can be combined into one prediction only. Simply combining those decisions in the case of classification is voting. This approach is used by both bagging and boosting. However, the individual models are derived by bagging and boosting in different ways. The same weights are taken by the models in bagging while weighting is given to more successful models in boosting as an executive may put alternative results on a variety of experts' advice relying on their previous correct estimations. The experts are individual decision trees which are made united by making them vote on every test. For a case that one gets more votes than other classes, it is considered as correct. When predictions are made by more number of votes, they are more reliable since there are more voters [24, 25].

3.2.4. Naïve Bayes (NBS)

Naive Bayes is one of the probabilistic approaches that utilize semantics in order to represent, use, and learn knowledge. The maximum aposterior (MAP) rule is as follows: an approximation for classifying a test sample *X* is to construct a probabilistic model to estimate the posterior probability $P(y \mid x)$ of the different *y*'s and to estimate the one with the greatest background probability [26]. In the following formula, Bayes theorem is represented in Equation (2):

$$P\left(y|x\right)=\frac{P(x\mid y)P\left(y\right)}{P\left(x\right)} \quad (2)$$

3.2.5. K-Nearest Neighbor (KNN)

The *K*-nearest neighbor classifier, which only stores the training set, is a lazy learning approach because there is no clear training process. It learns by analogy which means the comparison of a provided test tuple with training tuples which are similar. These tuples must be the closest ones to the unknown tuple. A distance metric like Euclidean distance describes the "closeness". In order to classify *k*-nearest neighbor, the tuple that is not known is selected as the most common class among its *k*-nearest neighbors. The rate of *k* can be determined experimentally [26].

3.2.6. All K-Edited Nearest Neighbors (AENN)

The efficiency of the classifiers designed under such circumstances which we generally want to optimize will depend heavily on the quality of the training data but also on the ruggedness of the classifier against noise. So training or test data with noise are complex problems and often difficult to achieve accurate solutions [27]. Noise in data can influence the intrinsic characteristics of a classification problem, as this can lead to new properties being introduced into the problem region. The dataset collected from the real-world are never flawless and often distorted that may inhibit the system efficiency. Therefore, data gathered from real-world problems are never perfect and often suffer from corruptions that may hinder the performance of the system. In order to have a clean data from all the movies classes, we employed All K-Edited Nearest Neighbors [28].

3.3. Performance Metrics

In literature, researcher have used different performance metrics to classify movie dataset. In this study, eight popular performance metrics such as accuracy (A_c), sensitivity (S_e), specificity (S_p), positive predicted value (PPV), negative predicted value (NPV), area under the curve (AUC), confidence interval (CI) and Kappa (K) are chosen. The mathematical representation of the measures are depicted in Equations 3-10.

$$A_C = 1 - \frac{F_N + F_P}{T_N + F_N + T_P + F_P} \tag{3}$$

$$S_e = \frac{T_P}{T_P + F_N} \tag{4}$$

$$S_p = \frac{T_N}{T_N + F_P} \tag{5}$$

$$PPV = \frac{T_P}{T_P + F_P} \tag{6}$$

$$NPV = \frac{T_N}{T_N + F_N} \tag{7}$$

$$AUC = \frac{1}{n_p n_n} \sum_{i=1}^{n_p} \sum_{j=1}^{n_n} I\left((D_p > D_n\right) + \frac{1}{2} I(D_p > D_n)) \tag{8}$$

$$CI = \mu \pm Z \frac{\sigma}{\sqrt{n}} \tag{9}$$

$$K = \frac{P_o - P_e}{1 - P_e} \tag{10}$$

Where A_c is the accuracy, S_e is the sensitivity, S_p is the specificity, PPV is the positive predicted value, NPV is the negative predicted value, AUC is the area under the curve, CI is the confidence interval, K is the Kappa, T_P is the true positive, T_N is the true negative, F_P is the false positive, F_N is the false negative, n_p is the total number of positive responses, n_n is the total number of negative responses, D_P is the distributions of the diagnostic variable in the negative responses, D_n is the distributions of the diagnostic variable in the positive responses, μ is the mean, Z-value from the table,σ is the standard deviation, n is the number of observation, P_o is the probability of the observed accuracy and P_e is the probability of expected accuracy obtained from the confusion matrix.

4. Results and Discussion

In this section, we provide the experimental results of our study on five machine learning used in this paper. Figure 1 display the correlation coefficients of duration, num_voted_users, num_user_for_review, title_year and imdb_scores. There is high correlation between num_voted_users and num_user_for_review while there is low correlation between imdb_scores and num_voted_users. Meanwhile, the imdb_scores have a negative correlation coefficient with title_year.

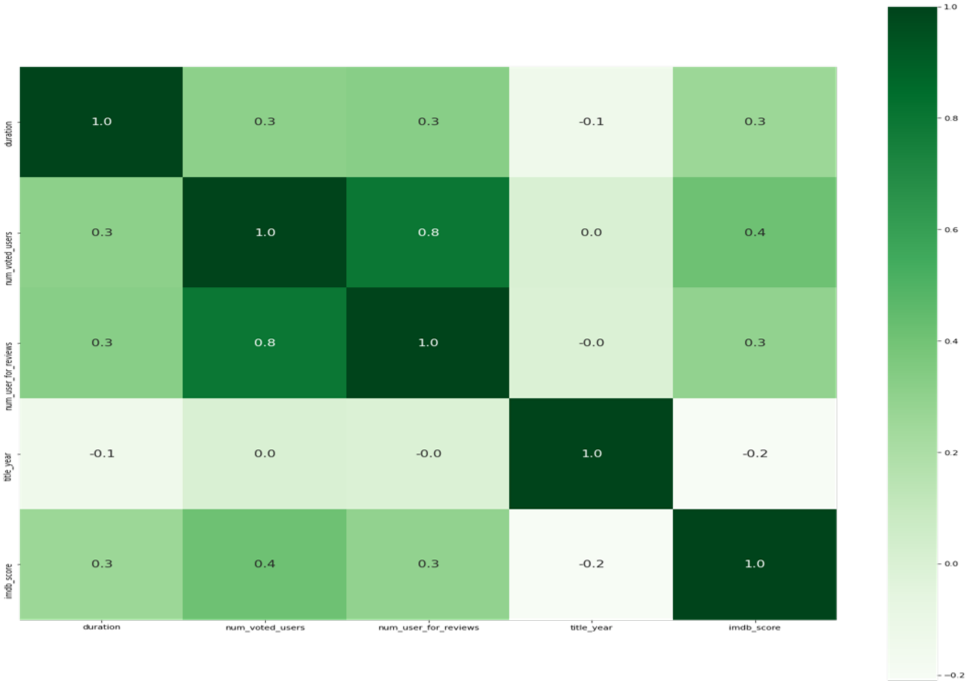

Figure 1 Correlation coefficients of movie dataset

Figure 2 displays top ten movie director obtained from the dataset. Steven Spielberg has the highest appearance than others with 14.28% followed by Woody Allen with 12.08%.

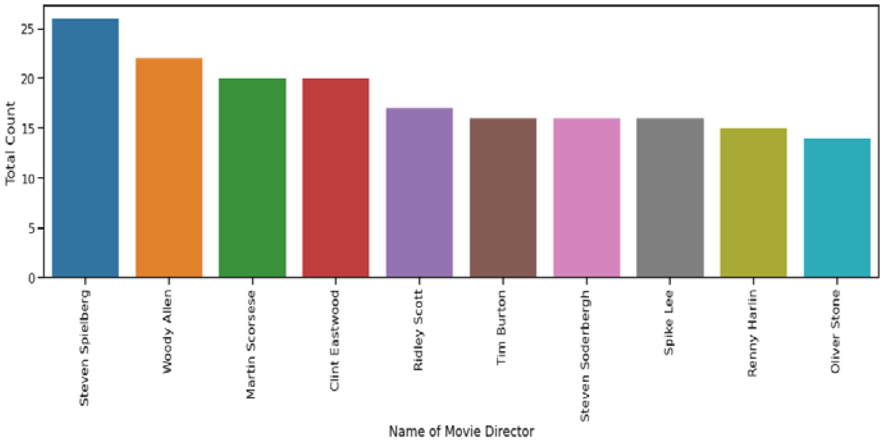

Figure 2 Bar plot of movie director

Figure 3 is the bar plot of top ten Movie director in x-axis while duration is in y-axis. Taylor Hackford has the highest duration in a movies followed by Michael Cimino.

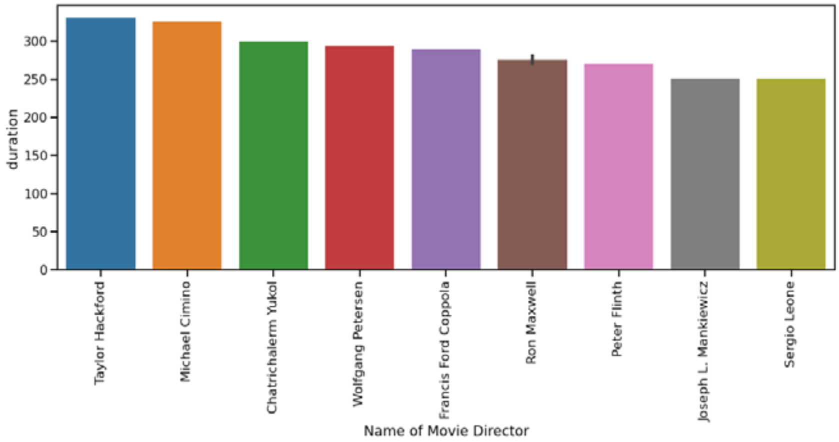

Figure 3 Barplot of Movie director vs duration

In Figure 4 the movie director with the highest vote is Frank Darabont followed by Christopher Nolan.

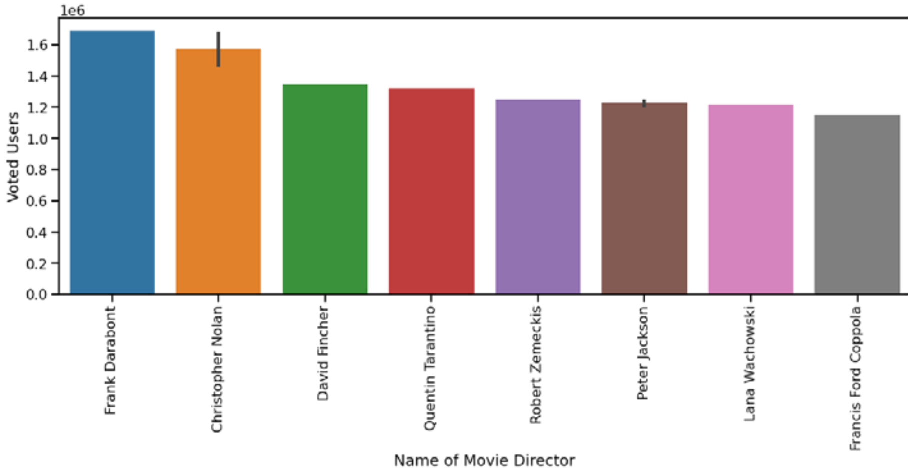

Figure 4 Bar plot of Movie director vs voted users

Figure 5 displays the movies with the highest language. English movies came first with 95.1% followed by French with 1.49%. This shows that people watched English movies than other movies.

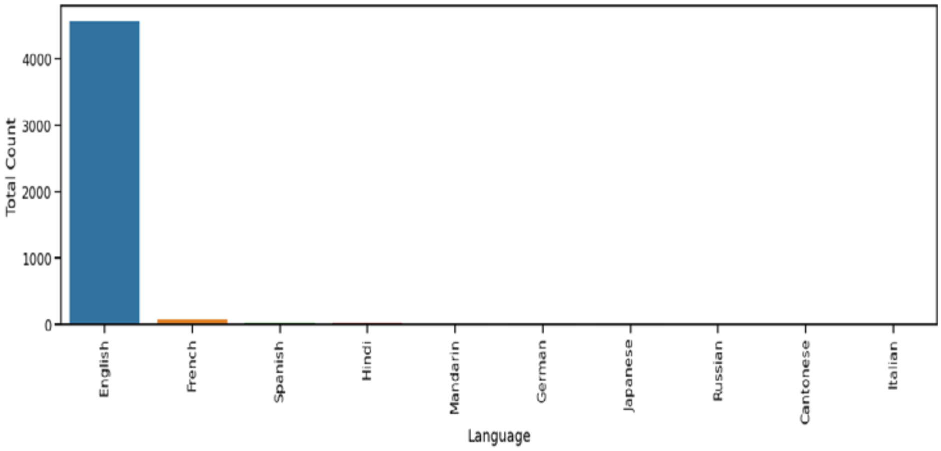

Figure 5 Bar plot of Movie Language

Figure 6 display Country with the highest movie production. USA came first followed by United Kingdom (UK) and France.

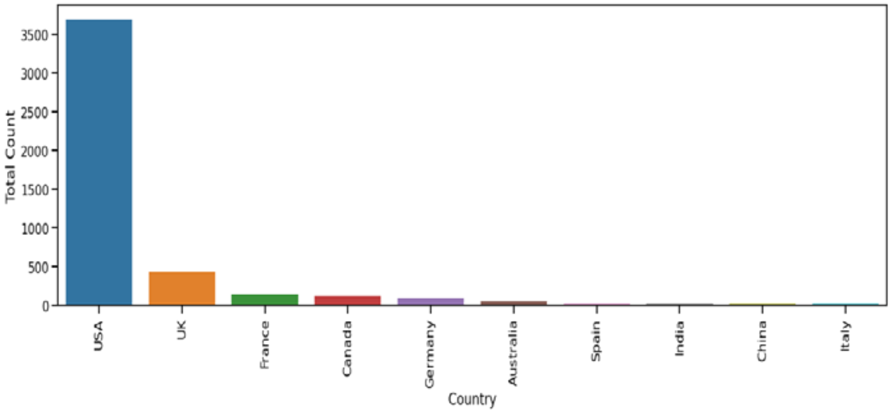

Figure 6 Bar plot of Country with movie production

Figure 7 shows that comedy has the highest number of viewers followed by action and drama. This shows that people prefer to watch comedy due to its ability to make us laugh, relieves stress, renew energy, and develops sense of humor and so on. The movie dataset consists of noise, to increase the quality of training movie dataset. Technique such as All K-Edited Nearest Neighbors (AENN) was used to filter noise from the dataset.

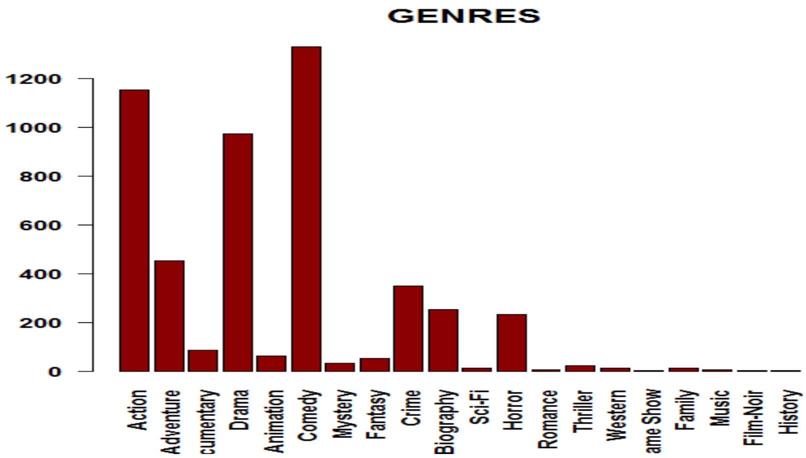

Figure 7 Genres Movies dataset

AENN filtered sci-fi, western, game show, family, music, film noir and history from the dataset as shown in Figure 8.

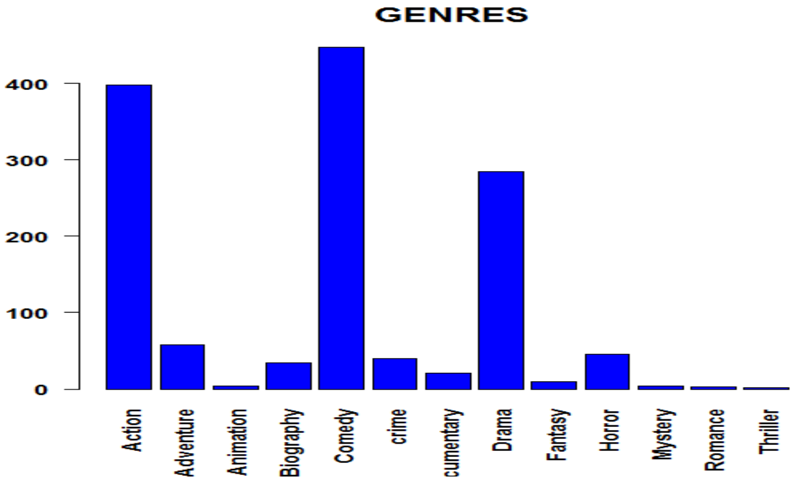

Figure 8 Genre Filtered movie dataset

Table 2 shows the performance metrics of filtered movies dataset. Different performance metrics were estimated including: sensitivity (S_e), specificity (S_p), positive predicted value (PPV), negative predicted value (NPV), overall accuracy (A_c), kappa (K), 95% confidence intervals (CI) and area under the ROC curve (AUC). MLR, SVM, NBS and KNN failed to classify movie dataset as shown in the performance metrics such as S_e, S_p, PPV, NPV. The sensitivity of the MLR, SVM, NBS and KNN is within the range of 0-80% as shown in Table 2. BAG on the other hand perform excellently well to classify movie dataset as shown in the performance metrics such as S_e, S_p, PPV, NPV. The sensitivity of the BAG is within the range of 99-100%. This shows that BAG performs better than MLR, SVM, NBS and KNN.

Table 2 Performance of machine learning movies dataset

Model	Genres	Sensitivity	Specificity	PPV	NPV
MLR	Action	63.82	86.36	66.15	85.11
	Adventure	1.72	100	100	95.77
	Animation	0.00	99.92	0.00	99.70
	Biography	14.70	99.69	55.55	97.83
	Comedy	79.19	71.57	57.94	87.43
	Crime	0.00	100	-	97.04
	Documentary	66.67	99.24	58.33	99.47
	Drama	61.62	87.35	56.45	89.53
	Fantasy	0.00	100	-	99.25
	Horror	6.52	99.61	37.50	96.79
	Mystery	0.00	100	-	99.70
	Romance	100	100	100	100

	Thriller	0.00	100	-	99.85
SVM	Action	66.58	93.70	81.54	87.04
	Adventure	15.55	100	100	96.34
	Animation	0.00	100	-	99.70
	Biography	11.76	100	100	97.77
	Comedy	89.71	69.91	59.58	93.22
	Crime	17.50	100	100	97.54
	Documentary	38.09	99.92	99.92	88.88
	Drama	67.25	90.16	90.16	64.53
	Fantasy	0.00	100	100	-
	Horror	39.13	99.46	72.00	97.88
	Mystery	0.00	100	-	99.70
	Romance	100	100	100	100
	Thriller	0.00	100	-	99.85
BAG	Action	99.75	99.69	99.25	99.89
	Adventure	96.55	100	100	99.84
	Animation	100	100	100	100
	Biography	100	100	100	100
	Comedy	100	99.67	99.33	100
	Crime	97.50	100	100	99.92
	Documentary	100	100	100	100
	Drama	99.30	100	100	99.81
	Fantasy	100	100	100	100
	Horror	100	100	100	100
	Mystery	100	100	100	100
	Romance	100	100	100	100
	Thriller	100	100	100	100
NBS	Action	61.06	82.06	58.70	83.46
	Adventure	10.34	99.76	66.66	96.12
	Animation	100	99.85	66.66	100
	Biography	11.76	100	100	97.77
	Comedy	81.21	70.58	57.71	88.37
	Crime	12.50	100	100	97.39
	Documentary	90.47	98.64	98.64	99.84
	Drama	44.01	95.68	95.68	86.52
	Fantasy	60.00	99.40	99.40	99.70
	Horror	26.08	99.54	99.54	97.44
	Mystery	50.00	100	100	99.85
	Romance	100	97.10	7.14	100
	Thriller	0.00	100	-	99.85
KNN	Action	67.84	77.54	55.79	85.24
	Adventure	12.06	99.07	36.84	96.17
	Animation	0.00	100	-	99.70
	Biography	0.00	100	-	97.48
	Comedy	72.04	69.47	53.85	83.40
	Crime	10.00	99.54	40.00	97.31
	Documentary	33.33	99.24	41.17	98.95
	Drama	33.09	90.15	47.23	83.50
	Fantasy	99.85	99.85	50.00	99.40
	Horror	17.39	99.23	44.44	97.14

Mystery	25.00	100	100	99.77
Romance	33.33	100	100	99.85
Thriller	0.00	100	-	99.85

Comparing the overall statistics of movie dataset from Table 3 BAG performs better with overall accuracy of 99.56%. This indicates that out of the 100 genres with different movie dataset that are tested, BAG will pick up to 99% of it. Figure 9 is the area under the curve (AUC) of movie dataset respectively.

Table 3 Overall Statistics of movies dataset

MODEL	Accuracy	Kappa	95% CI	AUC
MLR	59.88	44.10	(0.5721, 0.6251)	69.12
SVM	67.06	54.06	(0.6448, 0.6957)	73.19
BAG	99.56	99.41	(0.9904, 0.9984)	99.91
NBS	58.62	43.13	(0.5594, 0.6127)	70.52
KNN	53.00	34.31	(0.5030, 0.5569)	62.90

Figure 9 is the area under the curve (AUC) of movie dataset respectively.

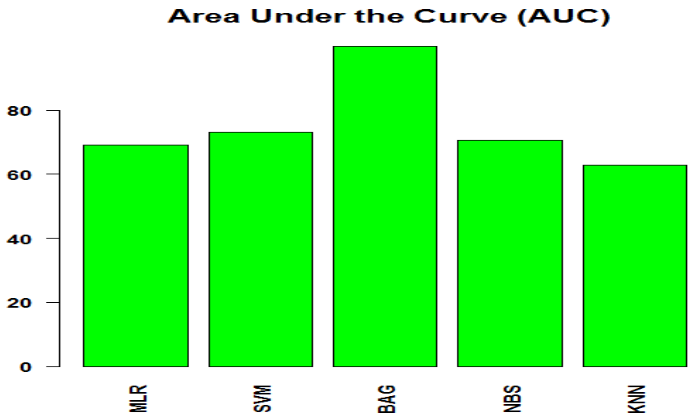

Figure 9 Area under the Curve (AUC) of the movie dataset

5. Conclusion

Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Five machine learning techniques were utilized in this study which consists of Multinomial logistic regression (MLR), support vector machine (SVM), Naïve Bayes (NBS), K-Nearest Neighbor (KNN) and Bagging (BAG) while All K-Edited Nearest Neighbors (AENN) were used to filter noise from movie dataset. BAG model has produced very impressive results for our predictive model.

The technique has proven to be very effective in classifying movie dataset. The performance of our model was assessed through several performance metrics such as sensitivity, specificity, PPV, NPV, overall accuracy, 95% confidence interval and AUC. BAG is able to achieve the accuracy 99.56%. Government and movies industry can also leverage on the technique due to its great influence to human culture and employment opportunities in the country.

Conflicts of 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] Latif, M.H., and Afzal, H. (2016). Prediction of movies popularity using machine learning techniques. *IJCSNS Int J Comput Sci Netw Secur* 16:127–131

[2] Masih, S., and Ihsan, I. (2019). Using academy awards to predict success of bollywood movies using machine learning algorithms. *Int J Adv Comput Sci Appl* 10:438–446

[3] Quader, N., Gani, M. O., and Chaki, D. (2017, December). Performance evaluation of seven machine learning classification techniques for movie box office success prediction. In *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)* (pp. 1-6). IEEE.

[4] Hafeez, E. (2012). Motion pictures as an agent of socialization: A comparative content analysis of demography of population on Indian Silver Screen and reported crime news in Pakistan (1976 to 2006). *Business Review*, *7*(2), 23-50.

[5] Lee, K., Park, J., Kim, I., and Choi, Y. (2018). Predicting movie success with machine learning techniques: ways to improve accuracy. *Inf Syst Front* 20:577–588. https://doi.org/10.1007/s10796-016-9689-z

[6] Im, D., & Nguyen, M. T. (2011). Predicting box-office success of movies in the US Market. *CS229, Stanford University, Fall.*

[7] Simonoff, J. S., & Sparrow, I. R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, *13*(3), 15-24. https://doi.org/10.1080/09332480.2000.10542216

[8] Latif, M. H., & Afzal, H. (2016). Prediction of movies popularity using machine learning techniques. *International Journal of Computer Science and Network Security (IJCSNS)*, *16*(8), 127.

[9] Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, *30*(2), 243-254. https://doi.org/10.1016/j.eswa.2005.07.018

[10] Cizmeci, B., & Ögüdücü, Ş. G. (2018, September). Predicting IMDb ratings of pre-release movies with factorization machines using social media. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)* (pp. 173-178). IEEE. https://doi.org/10.1109/ubmk.2018.8566661

[11] Tang, T. Y., Winoto, P., Guan, A., & Chen, G. (2018, February). "The Foreign

Language Effect" and Movie Recommendation: A Comparative Study of Sentiment Analysis of Movie Reviews in Chinese and English. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing* (pp. 79-84). https://doi.org/10.1145/3195106.3195130

[12] Wang, H., & Zhang, H. (2018, January). Movie genre preference prediction using machine learning for customer-based information. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 110-116). IEEE. https://doi.org/10.1109/CCWC.2018.8301647

[13] Gallaugher J (2008) Netflix case study: David becomes goliath. Gall com:1–16

[14] Jaiswal, S. R., and Sharma, D. (2017, November). Predicting success of bollywood movies using machine learning techniques. In *Proceedings of the 10th Annual ACM India Compute Conference* (pp. 121-124).

[15] Quader, N., Gani, M. O., Chaki, D., and Ali, M. H. (2017, December). A machine learning approach to predict movie box-office success. In *2017 20th International Conference of Computer and Information Technology (ICCIT)* (pp. 1-7). IEEE.

[16] Lee, K., Park, J., Kim, I., and Choi, Y. (2018). Predicting movie success with machine learning techniques: Ways to improve accuracy. *Information Systems Frontiers*, *20*(3), 577-588. **https://dx.doi.org/10.1007/s10796-016-9689-z**

[17] Marović, M., Mihoković, M., Mikša, M., Pribil, S., and Tus, A. (2011, May). Automatic movie ratings prediction using machine learning. In *2011 Proceedings of the 34th International Convention MIPRO* (pp. 1640-1645). IEEE.

[18] Jernbäcker, C., and Pojan, S. (2017). Predicting movie success using machine learning techniques. Master of Science, Computer Engineering Thesis, School of Computer Science and Communication, KTH.

[19] Bristi, W. R., Zaman, Z., and Sultana, N. (2019, July). Predicting IMDb Rating of Movies by Machine Learning Techniques. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.

[20] Meenakshi, K., Maragatham, G., Agarwal, N., and Ghosh, I. (2018, April). A Data mining Technique for Analyzing and Predicting the success of Movie. In *Journal of Physics: Conference Series* (Vol. 1000, No. 1, p. 012100). IOP Publishing.

[21] Lee, J. H., Kim, Y. J., and Cheong, Y. G. (2020, August). Predicting Quality and Popularity of a Movie From Plot Summary and Character Description Using Contextualized Word Embeddings. In *2020 IEEE Conference on Games (CoG)* (pp. 214-220). IEEE.

[22] Abidi, S. M. R., Xu, Y., Ni, J., Wang, X., and Zhang, W. (2020). Popularity prediction of movies: from statistical modeling to machine learning techniques. *Multimedia Tools and Applications*, 1-35.