FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 6 No. 6, December, 2022, pp 224 - 228
DOI: https://doi.org/10.33003/fjs-2022-0606-1117

COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES FOR MOVIE PREDICTION

[1]Ayokunle A. Omotunde, *[2]Ogunsanwo, Gbenga, O., [3]Adekola, Olubukola, [1]Izang Aaron, [3]Abel Samuel B

[1]Department of Information Technology, Babcock University, Ilishan Remo, Ogun State, Nigeria
[2]Computer Science Department, Tai Solarin University of Education, Ijagun, Ogun State, Nigeria
[3]Computer Science Department, Babcock University, Ilishan Remo, Ogun State, Nigeria

*Corresponding authors' email: ogunsanwogo@gmail.com

ABSTRACT
The rate at which movies are being produced is increasing at exponential rates and it has become pertinent to ascertain success rate since the investment that goes into these movie creation runs in millions of dollars. A number of data mining-based methods, ranging from Support Vector Machine (SVM) to logistic regression, have been proposed with a varying level of success with SVM showing the most promising results. This paper aims to carry out a comparative analysis of the performance of Gradient Boosting and SVM algorithms in optimizing the prediction of movie success. The study developed a framework for the research methodology; the dataset used contained 33 movie attributes and 838 entries. The dataset was cleaned with six attributes; features were identified and selected from the datasets using four methods. These methods include: Analysis of Variance (ANOVA), Lasso Regularization, Combination of Lasso Regularization and Random Forest (RF). Model Formulation were done using Support Vector Machine (SVM) and Gradient Boosting Algorithm and the performance evaluation of the developed predictive models was done using accuracy, precision and recall values. The results shows that the accuracy of the Gradient Boosting algorithm is around 100%, SVM-Linear is 86 %, SVM-Poly is 88%, SVM-RBF is 88% and SVM-Sigmoid is 72%. The study concluded that Gradient Boosting algorithm is more robust in predicting movie success. Also recommended that comparison should be done with different machine learning techniques.

Keywords: Comparative Analysis, Movie, Support Vector Machine (SVM), Gradient Boosting Algorithm

INTRODUCTION

The rate at which movies are being produced is increasing at exponential rates and it has become pertinent to ascertain success rate since the investment that goes into these movie creation runs in millions of dollars. This calls for proactiveness in determining the success or failure of a particular movie and establishing what factors affect the movie so as to know how to go about the promotion of the movie which entails huge advertising deals to sponsor it before it is release. In reality, before a movie is released, the success of the movie relied heavily on media hype, previews and pre-release marketing campaign, but this does not determine or translate to the movie's success when released. The problem found here is that most producers, directors or stakeholders end up spending millions on movie budget without knowing if the movie will be a success or a failure in monetary value and for viewers. Money is wasted on ticket purchases that do not give any value in terms of satisfaction. In recent times, however, a number of data mining-based method, ranging from Support Vector Machine (SVM) to logistic regression, have been proposed with a varying level of success with SVM showing the most promising results. With the backdrop, this work aims to carry out a comparative analysis of the performance of Gradient Boosting and SVM algorithms in optimizing the prediction of movie success. While the objectives are to: collect and preprocess dataset, select important features for predicting movie success, build movie success prediction models based on Gradient Boosting and SVM algorithms, evaluate the performance of the selected models in predicting movie success.

Section two reviewed closely related works and section three concentrated on the methodology for the study in which the model developed was conceptualized. Section four discussed the proof of conceptual model and results while section five stipulates the conclusion.

A number of researchers have worked on prediction of movies' success rate leveraging on different approaches such as news media, social media, and web media Saraee (2004) and Zheng & Skiena (2009) while Mestya'n & Yasseri (2013) predicted popularity of movies using articles from Wikipedia and takes movie information such as actors, director, genre and released date etc. from metacritic and financial data like budget, opening week gross revenue from the numbers. Logistic regression and support vector machine was used in predicting popularity of movies in Dwi *et al.,* (2019), Miryala *et al.,* (2017) used machine learning algorithm, which is imported from the spark and compared the efficacy and efficiency of the Alternating Least Squares (ALS) algorithm with Singular Value decomposition, K-Nearest Neighbor algorithm, and Normal predictor algorithm. The ALS algorithm is justified by theory and demonstrated on actual user data from Movie Lens. Mahesh *et al.,* (2010) leveraged on mean absolute error, Person's correlation coefficient and linear regression to show that review text can substitute for metadata and improve over it for prediction. Nithin *et al.,* (2014) Used IMDB data, rotten tomatoes and Wikipedia data about the movie and machine learning algorithms are applied on it like linear regression, SVM regression and logistics regression. Komal *et al.,* (2018) predicted movie success by a using historical data of actor, actress, music director, writer, director, marketing budget and release date of the new movie. Logistic regression, simple logistic, multilayer perception, J48, Naïve Bayes and PART were applied to dataset gotten from iMDB (Muhammad and Afzal, 2016). The highest accuracy was recorded with logistic regression and simple logistic.

METHODOLOGY

In section, the methodology applied to this research work is visibly stated. It starts with a description of the framework for the research methodology, which explains the series of steps

required: starting from the description of the dataset, data collection and pre-processing, model formulation and performance evaluation of the developed predictive models. In addition, the selected machine learning algorithms selected for the formulation of the predictive model were presented. The figure 1 shows the framework for the model.

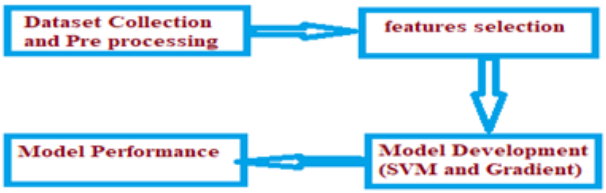

Figure 1: Model framework

Data Sources

A dataset of movies' attributes was sourced to conduct the study from Kaggle.com online repository database. The data set contained 33 movie attributes and 838 entries. The dataset was cleaned and six attributes such as 'Title', 'Genre', 'Description', 'Director', 'Actors', and 'Year' were removed from the dataset. This makes a total of 27 columns (attributes) and 838 rows. To construct the models, the dataset was divided into one dependent variable and 26 independent variables. Since the model aims to predict movie success. Success was selected as the response variable, while the remaining 26 variables were selected as the predictors. These include 'Rank', 'Runtime (Minutes)', 'Rating', 'Votes', 'Revenue (Millions)', 'Metascore', 'Action', 'Adventure', 'Animation', 'Biography', 'Comedy', 'Crime', 'Drama', 'Family', 'Fantasy', 'History', 'Horror', 'Music', 'Musical', 'Mystery', 'Romance', 'Sci-Fi', 'Sport', 'Thriller', 'War', and 'Western.

Feature Selection Methods

In order to improve model accuracy and construction time, important features were identified and selected from the datasets using four methods. These methods include:

i. Analysis of Variance (ANOVA)
ii. Lasso Regularization
iii. Combination of Lasso Regularization and ANOVA
iv. Random Forest (RF)

Model Development

In order to develop the model, two machine learning algorithms were selected for the study which includes: Support Vector Machine (SVM) and Gradient Boosting Algorithm

1) Support Vector Machine (SVM)

SVM classification models based on four karnels were used in the study. This includes the Linear, Poly, RBF and Sigmoid.

2) Gradient Boosting Algorithm

Using the Randomized SearchCV hyper parameter tuning method, a highly robust Gradient boosting model was established for the study. The following are the optimized parameters of the model. Best Model parameters: {'subsample': 0.75, 'n_estimators': 1000, 'min_samples_split': 2, 'min_samples_leaf': 3, 'max_features': 2, 'max_depth': 2, 'learning_rate': 0.15}. Using the optimized parameters, a model accuracy of 0.997 was achieved. Hence the selected parameters were used to develop the model used in the study. The figure 2 shows the k-fold validation in terms of AUC plot of the Gradient Boost model.

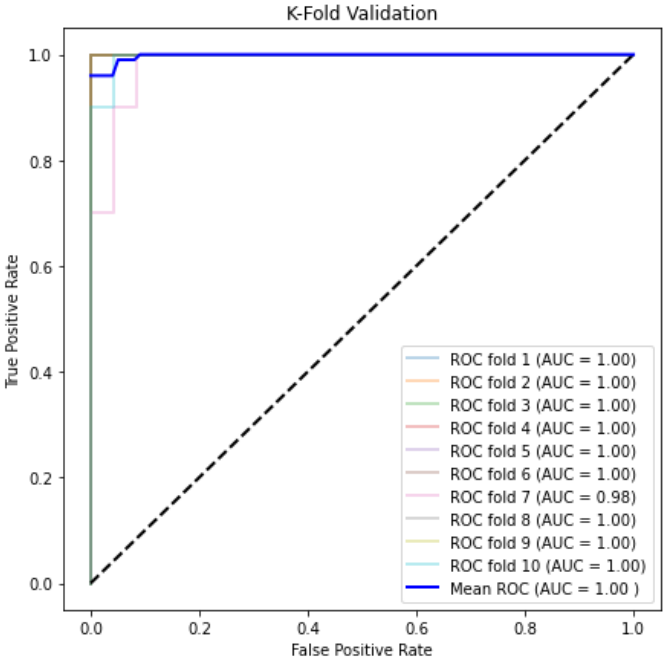

Figure 2: K-Fold Validation

RESULTS AND DISCUSSION

Important features selection

The figure 3 shows the most important feature for predicting the success of a movie. Based on the combination of Lasso Regularization and ANOVA, animation, Sci-Fi, Roman, Horror or Drama are more important for predicting a movie's success

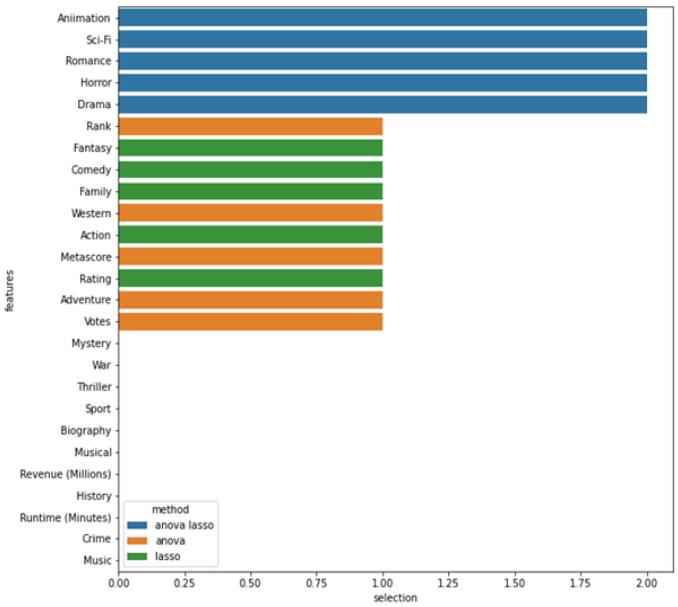

Figure 3: Lasso Regularization and ANOVA

The figure 4a, 4b shows the important feature for predicting movie success based on Random Forest algorithm. The figure shows that revenue (million), votes and rating are more important for predicting the success of a movie.

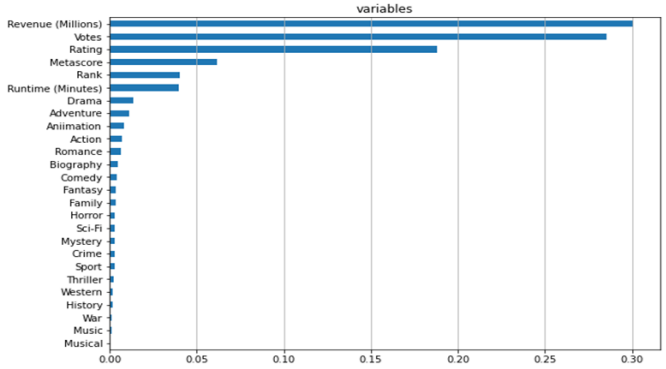

Figure 4a: Important features – variable

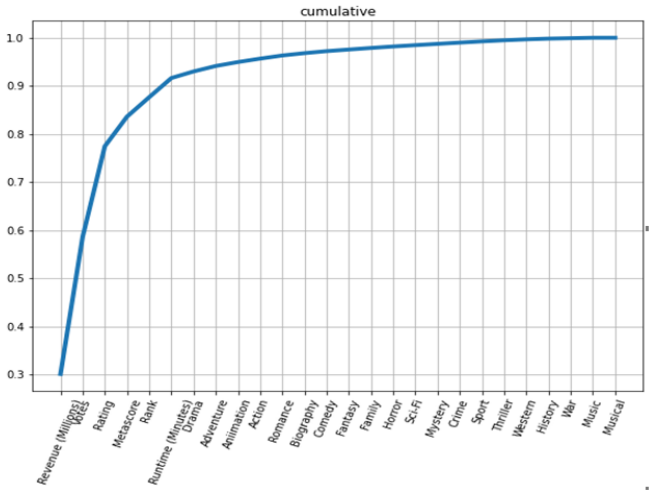

Figure 4b: Important features - cumulative

Based on the important feature selection algorithms, 24 features were selected for model development.

Model Performance

The Table 1 shows the selected model performance. Based on the table Gradient Boosting Model outperformed other models in terms of accuracy, precision and recall values.

Table 1: Summary of Model Performance

Machine Learning Models	Accuracy	Precision	Recall
SVM-Linear	0.86	0.68	0.47
SVM-Poly	0.88	1.0	0.32
SVM-RBF	0.88	0.85	0.39
SVM-Sigmoid	0.72	0.04	0.02
Gradient Boosting	1.0	1.0	1.0

The outcome of the study shows that the general accuracy of the Gradient Boosting model is around 100%. It predicted 100% of 1s (movie success) correctly with a precision of 100% and 100% of 0s (movie failure) with a precision of 100%, SVM-Linear is 86 % accuracy and 68% precision , SVM-Poly is 88% accuracy and 100% precision, SVM-RBF is 88% accuracy and 85% precision and SVM-Sigmoid is 72% accuracy and 4% precision.Also, based on the outcome of the model developed it can be deduced that Gradient Boosting algorithm and SVM-Poly were more consistent in the prediction. As shown in Table 1

In order to understand these metrics better, the results are broken down in a confusion matrix as shown in figure 5 and figure 6

```
Accuracy (overall correct predictions): 1.0
Auc: 1.0
Recall (all 1s predicted right): 1.0
Precision (confidence when predicting a 1): 1.0
Detail:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       208
           1       1.00      1.00      1.00        44

    accuracy                           1.00       252
   macro avg       1.00      1.00      1.00       252
weighted avg       1.00      1.00      1.00       252
```

Figure 5: Performance of Gradient Boosting

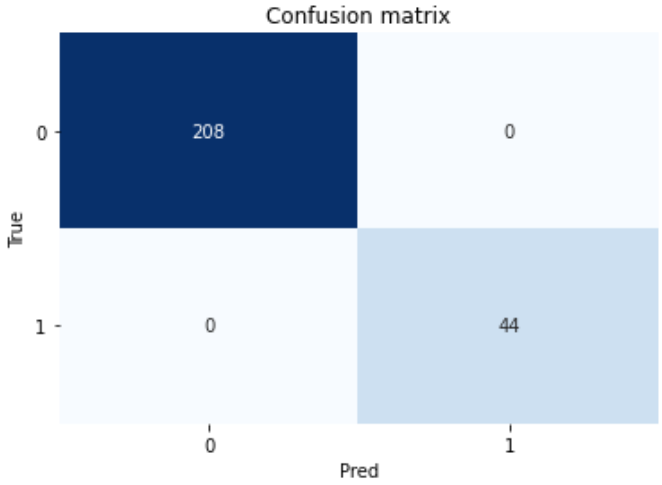

Figure 6: Confusion matrix of the gradient model.

CONCLUSION

Data mining techniques has a lot of prospects for predicting movies success so as to bring a total transformation in the entertainment industry. This technology if effectively used can assist to increase the profit and reduce the loss of the stakeholders in the movie industry. The study concluded that Gradient Boosting algorithm is more robust in predicting movie success. Based on the outcome of the model developed it can be deduced that Gradient Boosting algorithm and SVM-Poly were more consistent in the prediction. The study recommends that comparison should be done with different machine learning techniques

REFERENCES

Saraee, M., White, S., &Eccleston, J. (2004). A data mining approach to analysis and prediction of movie ratings. In A.

Zanasi, N. F. Ebecken, & C. A. Brebbia (Eds.), Data Mining V: Data Mining, Text mining and their Business Applications. UK: WIT Press / Computational Mechanics.

Zheng, W., &Skiena, S. (2009). Improving Movie Gross Prediction through News Analysis. International Conference on Web Intelligence and Intelligent Technology (pp. 301 - 304). Department of Computer Science Stoney Brook University.

Mestya'n, M., &Yasseri, T. K. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLoS ONE. doi:e71226.doi:10.1371/journal.pone.0071226

Dwi, R., Rosyida, I., &dan-Santi, W. P. (2019). Predicting Popularity of Movie using Support Vector Machines. INFERENSI, 2(1), 13 -17.

Miryala, G., Gomes, R., & Dayananda, K. R. (2017). Comparative Analysis of movie recommendation System using collaborative filtering in Spark Engine. Journal of Global Research in Computer Science, 8(1).

Mahesh, J., Dipanjan, D., Kevin, G., & Noah, A. S. (2010). Movie Reviews and Revenues: An Experiment in Text Regression. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (pp. 293 - 296). Los Angeles: Association for Computational Linguistics.

Nithin, V. R., Pranav, M., Sarath, B., &Lijiya, P. B. (2014). A Predicting Movie success based on IMDB data. International Journal of data mining and techniques, 243-254.

Komal, G., Dhiral, S., Nirav, W., Mitul, S., &Ramanand, Y. (2018). Movie Success Prediction. IOSR Journla of Engineering, 6, 66-69.

Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (2018). A Data mining Technique for analysing and predicting the success of a movie. National Conference on Mathematical Techniques and its Applications (NCMTA18).

Muhammad, L., & Afzal, H. (2016). Predicting of Movies Popularity Using Machine Learning. International Journal of Computer Science and Network Security, 16(8), 127-131