# A Comprehensive Study of Various Statistical Techniques for Prediction of Movie Success

*Agarwal Manav[1], Venugopal Shreya[2], Kashyap Rishab[3]
Department of CSE
PES University
Bangalore, India
[1*]manav.ag.3052@gmail.com
[2]shrey1010.svg@gmail.com
[3]rishabkashyap14@gmail.com

**Abstract:** The following paper summarizes the use of various types of machine learning models such as regressions, neural networks and classifiers, as well as a time series model used to predict the success rate of a given movie. We know the importance of a movie being a hit, or successful, hence our paper describes various models and studies their individual outcomes. We thereby create a comparative study between the various properties of each model, such as the accuracy, the feasibility, performance and most importantly, its statistical significance. Through this study, we aim to come to a conclusion on the usability of each of these models and the importance they play in the prediction of any domain of data. Choosing Movie Success prediction we have an accuracy of about 88% which is at power with the current state of the art models.

**Keywords:** Machine Learning Models, Time Series, Comparative Study, Movie Success, Neural Network, Statistical Significance.

## 1 Introduction

One of the most important contributing factors to the entertainment industry are movies, which turns out to be one of the highest revenue-generating businesses from a business perspective[4]. A majority of the population love to watch a variety of movies, and their choices are determined based on the various factors that contribute to the type of movie such as the genre of the movie. Most of the people thus look into the ratings of a given movie before they proceed to watch it to identify it as a movie worth watching for them. These ratings come from a variety of sources, some of which include popular websites such as Rotten Tomatoes, IMDb and many more. Thus, our analysis involves the study of these user ratings as well as the other factors that affect the movie and this enables us to predict whether a movie is truly a successful one or not.

## 2 Literature Survey

The inspiration of movie success predictions comes from the dataset itself that we retrieve from Kaggle[1] which explains various attributes related to a number of movies. We also perform web scraping[7] from the IMDb official site to extract all the movies released in 2020 and during the pandemic. As mentioned before, movies play a vital role in contributing to the entertainment industry[4]. Since a lot of people get reviews and ratings from this site, it is important to sum up the various attributes and try predicting the best possible outcome of the movie's success. Many papers have made use of various Machine Learning models such as K-Means[5], SVM [18], and various Regression techniques to predict the closest and most accurate ratings for a given movie.

## 3 Proposed Methodology

Similar to the methods mentioned above, through our paper we perform similar analytics on the IMDb dataset and create a comparison between the various models used [10][16], and try to pinpoint the most effective model in the process. The figure given describes the various steps involved in doing so.

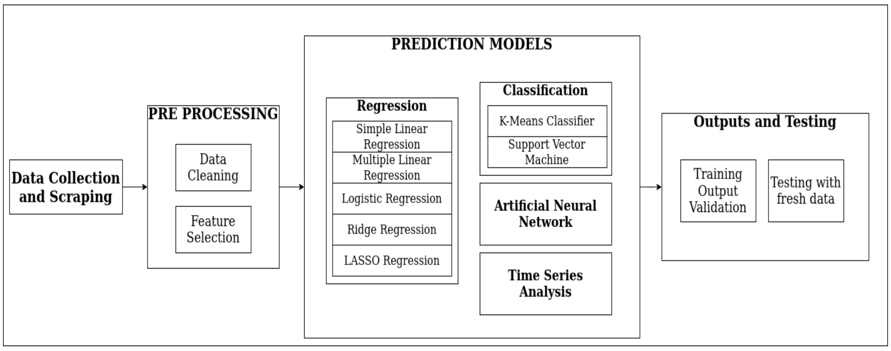

**Fig. 1.** Proposed Methodology Flowchart

## 4 Dataset Description

The dataset that we have considered consists of a list of movies which consists of 81274 movies by title[1]. The dataset is divided into 4 files. In this case, we look into only 2 of the files. The first file we look into consists of all the movie titles and the attributes related directly to the movie, such as the duration, the genre, top voters in both the critic as well as user perspectives, and finally the total votes for each movie. The second dataset looks into the average votes given for each movie as well as the number of votes given by people within a given age group. We combine the two datasets which contain predominantly numeric attributes. We additionally extract a dataset consisting of all the movies released in 2020 by scraping it from the official IMDb website. We perform the same set of steps as described above to this dataset, and we treat it as a purely testing dataset for all the models that follow.

## 5 Pre-Processing

Traditional pre-processing methods such as removing null values, outliers and other basic steps have been applied. Apart from these, another pre-processing step used is the MultiLabelBinarizer which splits the genre attribute in particular into numerical binary values for each movie. If the movie has a certain genre, then the binarizer labels the value as 1 and if it doesn't it labels it as 0. We hence use the same to convert the entire categorical attribute into a set of numeric attributes. We have taken all movies from 1990 to 2015 as the training dataset and all movies after 2015 as our validation dataset. The dataset contains an attribute called the metascore which ranges from 0 to 100 and is used as a measure to depict the success of a movie. A higher metascore implies that the movie was more successful. Hence this is used as the dependent variable for the complete study. To classify a movie as a success or a failure, the metascore value is divided into classes or bins, representing movies that are a hit or a flop or mediocre [6]. The predicted values can be classified under this partitioning, and the accuracy is based on how well it is sorted.

## 6 Regression Methods

Regression is a technique that uses the Ordinary Least Squares method to predict certain values given certain inputs. The various methods we use here use almost the same attributes as we have taken before, or a subset of these attributes. The metascore value predicted is thus compared to the true value to get the accuracy, residual, and other statistics that we prove through our analysis.

### 6.1 Simple Linear Regression

As we all know, SLR is the most basic regression form. Since it involves only one independent variable against the metascore the most suitable attribute is selected. To do so, we use the Variance Inflation Factor (VIF) method to show the attributes with their multicollinearity with respect to each other. The highest value is considered to be the best prediction value, and in this case, we obtain the attribute to be the top1000_voters_ratings. We hence use this attribute in the SLR model and train the model with the existing values.

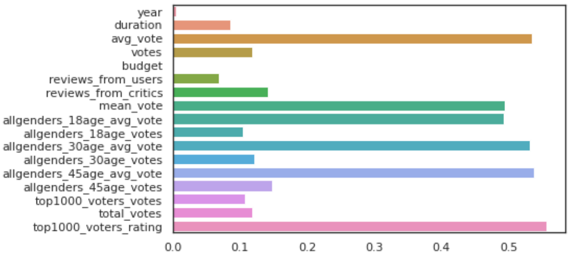

**Fig. 2.** Diagram to show the VIF values for all attributes

Once we have our trained model, we use the testing dataset to predict the future values, and then make a comparison between the true and predicted values. The accuracy of the model is calculated using a confusion matrix. On completion of this, we proceed to perform tests on the model to prove its statistical significance.

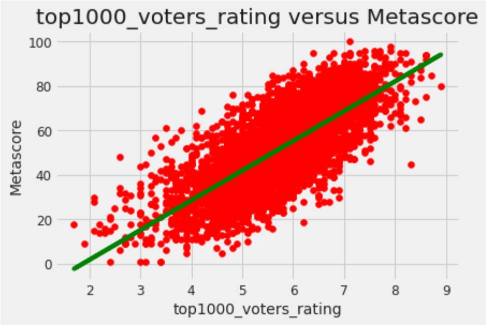

**Fig. 3.** SLR model

We then test the same on our 2020 dataset, taking the avg_vote attribute instead of the top100_voters_ratings attribute in the previously trained model and get an accuracy of 0.5833. The model has already been tested for its significance and hence we do not test it again for the 2020 dataset.

### 6.2 Multiple Linear Regression

Multiple Linear regression (MLR) performs a process almost identical to SLR, but with multiple independent variables used to predict the same target variable.

Similar to the procedure above, the target variable is predicted using the top1000_voters_rating attribute along with fifteen other attributes. Here, twelve of these attributes come under the MultiLabelBinarized values of the genre. The VIF test performed on them shows us the initial attributes that can be considered to plot the model. Once again we perform the OLS tests and several hypothesis tests for attribute selection to show the significance of the values. The model that we get gives us an accuracy of 0.7116**.** An example of the MLR model in three dimensions using two independent variables is shown below. We perform the same analysis using our 2020 dataset, taking the attributes avg_vote and duration for plotting our MLR model, and result with an accuracy of 0.608.

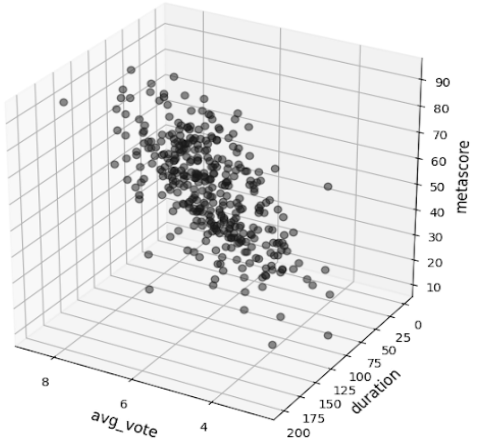

**Fig. 4.** MLR model

## 6.3 Logistic Regression

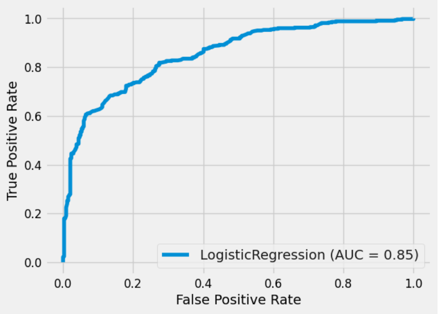

**Fig. 5.** *ROC-AUC curve for the logistic regression model*

Logistic Regression is a non-linear regression and a statistical technique for finding the existence of a relationship between a qualitative and a quantitative dependent variable and several independent variables or explanatory variables[3]. The deviation we take from our originally defined steps here is the fact that we divide our hit-flop classification into two domains, a successful or an unsuccessful movie based on the metascore value. This model results in an accuracy of 0.76. We plot an ROC-AUC curve using the confusion matrix that we have obtained. The same way we perform the Logistic analysis on our 2020 dataset to get an accuracy of 0.6833.

### 6.4 Regularization Techniques

Regularization is required to penalize certain features, and we have two regression methods for the same, namely the Ridge, and the LASSO regression models. Ridge regression is a technique used when the data suffers from high multicollinearity[2]. It uses the L2 regularization method to perform this process. The accuracy we get from this model comes upto 0.74. Plotting the same using our 2020 dataset we thus get an accuracy of 0.61.

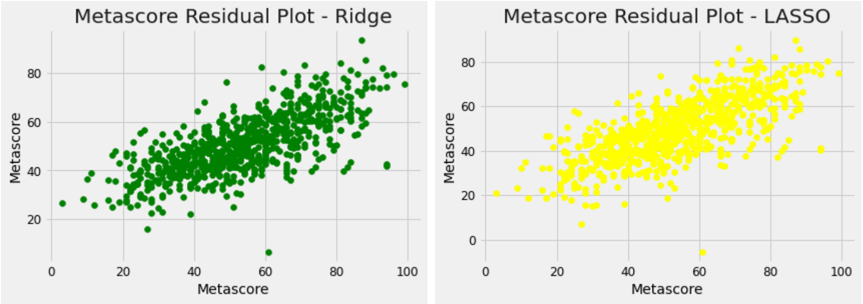

**Fig. 6.** Ridge and LASSO regression model residual plot

As we did for Ridge, we follow the exact same procedure in LASSO regression, which along with L2 regularization, uses central tendency to penalize. For the training and testing process, calculating the statistical values as well as obtaining an accuracy of 0.72. Plotting the same using our 2020 dataset we thus get an accuracy of 0.59.

## 7 Classification Methods

The classification models that we have used follow the unsupervised learning algorithm, rather they use self learning techniques. The two classification models that we use here are the Support Vector Machine and the K-Means classifier.

### 7.1 K-Means Classifier

Using the K-Means classifier would involve making the value of K as 3 in this case for each of the metascore bins[2]. The accuracy of the K-Means model is 0.5. Statistical tests for significance are performed. We plot the same for our 2020 dataset using our previously trained model and hence get an accuracy of 0.42.

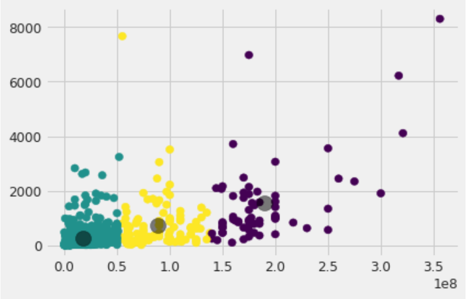

**Fig. 7.** K-Means plot to show the final centroid positions

### 7.2 Support Vector Machine

Instead of the conventional binary Support Vector Machine(SVM) the one that classifies into 3 categories with more than one hyperplane of separation is used. We can see that the model thus results in an accuracy of 0.71. The trained model is then used against the 2020 testing dataset to calculate the same and thus results in an accuracy of 0.62.

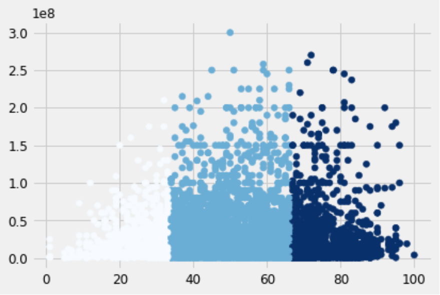

**Fig. 8.** SVM models for our initial dataset

## 8 Time Series Analysis

We begin by defining the method of Forecasting, which is one of the most important and frequently used applications in analytics, which focuses on the prediction of future values based on the present and past values. We say that the variable is forecasted into future values to perform analysis on patterns such as trend, seasonality and so on. We make use of the SARIMAX model for our analysis. We first establish all the necessary parameters to plot our time series graphs. We then initialize the various attributes that will be used to determine the forecasted values for metascore. Once we initialize these, we set the common index to be the date of release of the movie, which is the variable we use to depict the time series.

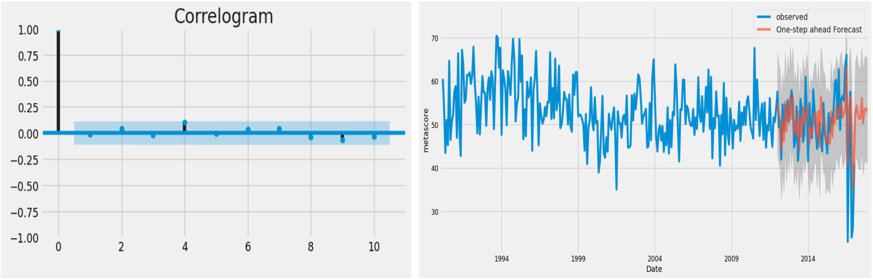

**Fig. 9.** Correlogram plot and Residual plot with Forecasted values

We then move on to plot our metascore value against the date_published attribute to get the time-series graph for the entire dataset, then plot the decomposition of the data into the trend, seasonality as well as the residual variations. To prove the stationarity of the data we use the Augmented Dickey-Fuller Test. The presence of seasonality in the plot leads us to use the SARIMAX model. Initially we define the function to execute all possibilities for SARIMAX from which one model is selected based on its AIC value. We can see that the model gives us a roughly accurate result for the forecast. We plot the same for our 2020 dataset, using the SARIMAX model we used before to get similar outputs.

## 9 Artificial Neural Network

In the Artificial Neural Network model we have proposed, we give the inputs based on the statistical significance they hold from the tests we have performed in the previous cases. On plotting the ROC curve as shown in the below image, we obtain a final result of 86.16% accuracy from the model as well as minimal loss incurred through the process. The loss is also displayed in the next image shown below. We have also tested this with the 2020 dataset obtained resulting in a 88.056 % accuracy.

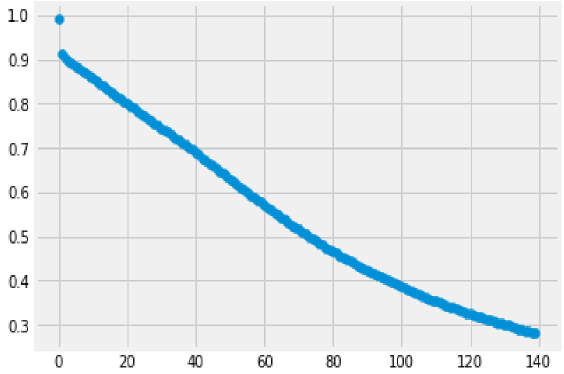

**Fig. 10.** Loss Curve for the ANN

# 10 Results and Interpretation of Values

## 10.1 Regression

Wald's test for Logistic Regression proved the statistical significance of the model. Hence the analysis shifts to the other regression models that are compared in Table 1. Since all the Durbin-Watson Test values are near 2 it can be safely said that there is no autocorrelation which gives these regression models a validity. In terms of R2 values and its variants Multiple Linear Regression seems to be showing the highest value. This is supported by the F statistic being high unlike that for Ridge and Lasso Regression indicating that the results are not significant. However the F-statistic value is significantly higher for Simple Linear Regression which seems to give it also more credit in the ranking of models as its R2 value was only marginally lower. In terms of normal distribution of errors all regressions except SLR are statistically significant with respect to their Jarque-Bera and Lagrange Multiplier values. Therefore it can be interpreted that out of these regressions only MLR and Logistic Regression can go for further analysis since they are statistically significant for all parameters.

**Table 1.** Regression

| X | Simple Linear | Multiple Linear | Ridge | LASSO |
|---|---|---|---|---|
| R2/ Pseudo R2 | 0.556 | 0.619 | 0.4855 | 0.46 |
| Adjusted R-Square | 0.556 | 0.618 | 0.48 | 0.4553 |
| F-Statistics | 6007 | 485.0 | 2.449e-37 | 2.477e-27 |

| | | | | |
|---|---|---|---|---|
| Durbin-Watson Test | 1.952 | 1.963 | 1.79127 | 1.7163 |
| Jarque-Bera (JB) Test | 9.617 | 0.91 | 34.778 | 90.18286 |
| Lagrange Multiplier Statistic | 14.237 | 145.0 | 208.301 | 185.15 |
| Accuracy | 0.71 | 0.711 | 0.72 | 0.72 |

## 10.2 Classification

The parameters have been tuned to obtain maximum silhouette score and based on accuracy it is seen that K-Means is not as accurate as SVM.

**Table 2.** Classification

| X | K-Means | SVM |
|---|---|---|
| Silhouette Test | 0.7021 | 0.13387 |
| Accuracy | 0.49934 | 0.71 |

## 10.3 Time Series Analysis

From the autocorrelation and partial autocorrelation plot results it is evident that the AR and MA parameters of the Time Series to be considered should be 1 each. This is supported by the Durbin Watson Statistic which indicates positive autocorrelation. The Augmented Dickey Fuller Test was showing stationarity based on the value given. The presence of seasonality was confirmed when the model with seasonality was giving a lower AIC value. The presence of exogenous variables were confirmed when they resulted in a reduction in the RMS value. The model was tuned to get the lowest possible Likelihood, AIC, BIC and HQIC. The Ljung Box Statistic leading to a p-value greater than the significance level also shows the validity of the model. The Jarque-Bera Statistic confirms the heteroscedasticity.

**Table 3.** Time Series Analysis

| X | Older movies analysis |
|---|---|
| Autocorrelation | Cuts of to 0 after 1 lag |
| Partial Autocorrelation | Cuts of to 0 after 1 lag |
| Augmented Dickey-Fuller Test Statistic | -10.462528698062133 |
| RMSE | 62.19 |
| Log-Likelihood | -965.539 |
| AIC | 1949.079 |
| BIC | 1982.679 |
| HQIC | 1962.512 |
| Ljung-Box(Q) | 24.64 |

| | |
|---|---|
| Skewness | -0.26 |
| Kurtosis | 4.13 |
| Jarque-Bera(JB) Test | 19.9 |
| Heteroscedasticity | 0.95 |
| Durbin-Watson Test | 1.4928194722663908 |

### 10.4 Artificial Neural Network

The neural network shows its superiority by giving a very high accuracy with the minimum possible loss. The tuned hyper-parameters are given below. The model does not show overfitting.

**Table 4.** Artificial Neural Network

| | |
|---|---|
| Attributes | 'duration', 'avg_vote', 'Action', 'Adventure', 'Animation', 'Biography', 'Comedy', 'Crime', 'Drama', 'Family', 'Fantasy', 'Horror', 'Mystery', 'Thriller' |
| Type | Multi-layer Perceptron Classifier |
| Architecture | İnput Layer: 14<br>Hidden Layer: 100<br>Output Layer: 3 |
| Output Type | Ternary |
| Initial Loss | 0.6915657421532461 |
| Final Loss | 0.1624720046108523 |
| Activation Function | Logistic |
| Optimizer | Adam |
| Early Stopping | True |
| Validation Fraction | 0.1 |
| Number of training examples | 4796 |
| Loss Curve Type | Strictly Decreasing |
| Testing results with 2020 dataset | 93.055555556 % |

### 10.5 Comparison of Valid Models with Similar Accuracy

SLR and Ridge and Lasso Regression were not considered as they proved to be invalid by the normality test and F Statistic respectively. Now 3 models of comparable accuracy exist as shown in Table 5. Hence, a Jaccard Index is used to find the similarity between the attributes obtained from the dataset and that scraped from the IMDb website. The Jaccard Index is used as an indication of the degree to which attributes used in the model are available. From this it can be concluded that SVM is the better model in terms of availability among the 3 followed by Multiple Linear Regression and Logistic Regression respectively.

**Table 5.** Comparison of Models with similar accuracy

| Model Name | Attributes | Attributes 2020 | Jaccard Index |
|---|---|---|---|
| Multiple Linear regression | ‘budget’,‘reviews_from_users’,‘review_from_critics’, ‘top1000_voters_ratings’, ‘Action’,‘Animation’,‘Crime’, ‘Drama’,‘Family’,‘Fantasy’, ‘Horror’, ‘Music’, ‘Musical’, ‘Mystery’,‘Sport’,‘Thriller’ | ‘duration’,‘Action’, ‘Animation’,‘Biography’,‘Drama’, ‘Horror’ | 4/18=0.222 |
| Logistic regression | ‘top1000_voters_rating’, ‘Action’,‘Crime’,‘Drama’, ‘Fantasy’,‘Mystery’,‘Romance’, ‘Sport’, ‘Thriller,’ ‘War’ | ‘avg_vote’,‘Action’ ,‘Crime’,‘Fantasy’,‘ Mystery’ | 4/11=0.18 |
| SVM | 'top1000_voters_rating', 'Action', 'Crime', 'Drama', 'Fantasy', 'Mystery','Romance', 'Sport','Thriller', 'War' | 'avg_vote', 'Action','Crime', 'Drama','Fantasy','Mystery','Thriller' | 6/11=0.545 |

## 11 Discussion

Our analysis on the prediction of a movie success using traditional regression methods was inspired by the fact that there were plenty of papers that used various machine learning models to predict the success of a movie. Given that these returned accurate results, we wanted to explore these techniques and the statistical significance of using such models. The tests were aimed at checking for the basic assumptions that follow the application of a model such as the normal distribution of errors for regression and so on. The reason for doing so was that we did not want any inconsistency or overfitting to affect our predictions. The end goal was to make the predictor useful in the real world. To corroborate this, the most recent 2020 data was scraped. This gave a sense of the attributes that will be available on immediate release of a movie.

## 12 Conclusion and Future Work

The prediction values from each of the models thus varies depending on their structural build as well as the method they use for prediction, and analysing these changes through comparison as well as statistical tests to prove its significance. By doing so it is seen that the Artificial Neural Network is the best model for prediction followed by the Support Vector Machine, Multiple Linear Regression and Logistic Regression that give comparable performance in terms of accuracy Logistic Regression being slightly higher. The availability of the attributes of these 3 models have also been analyzed where the Support Vector Machine has better availability. Following these 3 models, comes the K-Means with a much lower accuracy. Simple Linear Regression and the Regularization techniques are deemed invalid due to statistically insignificant results. Further, a Time Series Analysis that uses a SARIMAX model forecasts the metascore value effectively. From the models it is evident that some attributes like the top1000_voters_rating and the genres of the movie play a major role in the prediction of movie success. Also some genres have more predictability than others as is evident from the analysis shown previously. Our highest accuracy of 88.056% shows that it is a commensurating with the works of Abidi et al.[4] and Verma and Verma[6]. In the near future we aim to look into the various improvements that can be performed on the individual models to boost their performance, and hence broaden our perspective on the classification of a variety of other models with comparable performances to show the significance of each one of them.

## References and Bibliography

1. IMDb extensive dataset, Kaggle
2. U Dinesh Kumar, Wiley (2017), *Business Analytics: The Science of Data-Driven Decision Making*
3. William Navidi, McGraw Hill (2011), *Statistics for Engineers and Scientists:Indian Edition*
4. Popularity prediction of movies: from statistical modeling to machine learning techniques (Abidi et al. 2020)
5. S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74 (Na et al. 2010)
6. G. Verma and H. Verma, "Predicting Bollywood Movies Success Using Machine Learning Technique," 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 2019, pp. 102-105, doi: 10.1109/AICAI.2019.8701239. (Verma and Verma 2019)
7. J. Ahmad, P. Duraisamy, A. Yousef and B. Buckles, "Movie success prediction using data mining," *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Delhi, 2017, pp. 1-4, doi: 10.1109/ICCCNT.2017.8204173. (Ahmad et al. 2017)
8. Ericson, J., & Grodman, J. (2013). A predictor for movie success. *CS229, Stanford University*.
9. Early Predictions of Movie Success: the Who, What, and When of Profitability (Lash and Zhao 2016)
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)*, Jalandhar, India, 2018, pp. 385-390, doi: 10.1109/ICSCCC.2018.8703320 (Dhir and Raj 2018)

11. N. Darapaneni *et al*., "Movie Success Prediction Using ML," *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York City, NY, 2020, pp. 0869-0874, doi: 10.1109/ UEMCON51285.2020.9298145. (Darapaneni et al. 2020)
12. T. Sharma, R. Dichwalkar, S. Milkhe and K. Gawande, "Movie Buzz - Movie Success Prediction System Using Machine Learning Model," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 2020, pp. 111-118, doi: 10.1109/ICISS49785.2020.9316087.
13. Kumar, Saurabh. (2019). Movie Success Prediction using Data Mining For Data Mining and Business Intelligence(ITA5007) of Master of Computer Application School Of Information Technology and Engineering.
14. Lee, K, Park, J, Kim, I & Choi, Y 2016, 'Predicting movie success with machine learning techniques: ways to improve accuracy' Information Systems Frontiers, vol (in press). DOI: 10.1007/s10796-016-9689-z (Lee et al. 2018)
15. N. Quader, M. O. Gani, D. Chaki and M. H. Ali, "A machine learning approach to predict movie box-office success," *2017 20th International Conference of Computer and Information Technology (ICCIT)*, Dhaka, 2017, pp. 1-7, doi: 10.1109/ICCITECHN.2017.8281839. (Quader et al. 2017)
16. N. Quader, M. O. Gani and D. Chaki, "Performance evaluation of seven machine learning classification techniques for movie box office success prediction," 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), Khulna, 2017, pp. 1-6, doi: 10.1109/EICT.2017.8275242. (Quader et al. 2017; Quader et al. 2017)
17. W. R. Bristi, Z. Zaman and N. Sultana, "Predicting IMDb Rating of Movies by Machine Learning Techniques," *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944604.(Bristi et al. 2019)
18. V. Subramaniyaswamy, M. V. 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)*, Palladam, 2017, pp. 182-186, doi: 10.1109/ISS1.2017.8389394.(Bristi et al. 2019; Subramaniyaswamy et al. 2017)
19. Vr, Nithin & Pb, Sarath. (2014). Predicting Movie Success Based on IMDB Data. (Bristi et al. 2019; Subramaniyaswamy et al. 2017; Nithin et al. 2014**)**