Movie Success Rate Prediction Using Robust Classifier

Balaganesh N, Bhuvaneswari M S

Abstract – Film industry is a multi-billion-dollar industry where each movie earns over billions of dollar. Predicting the success of the movie is a difficult task because the success rate is influenced by various factors like running time, actor, actress, genre etc. In this paper a detailed study of machine learning algorithms such as Adaboost, SVM, and K-Nearest Neighbours (KNN) were done and was implemented on IMDB dataset for predicting box office. Based on the results, Adaboost classifier gives better performance compared to SVM and KNN classifier algorithms.

Keywords: Classifier, Movie Prediction, IMDb, Machine Learning.

I. INTRODUCTION

The passion of youngsters towards movies and the fear of producers and distributors whether to buy a movie or not is playing a major role towards the research in movie success rate prediction. It is more difficult for anyone to predict whether the movie will become a box office or Flop. The Producer and distributor needs to do a lot of research about the cast and crew and also about the story of the film in order to gain the minimum confidence for funding or buying the movie.

Many researches have been carried out to predict the success of before and after release of the movie. Since the success can be easily predicted after the release of the movie the accuracy of the prediction is high compared to prediction before release date. The success rate prediction can also be done based on the data derived from the like-minded viewers of the movie. The movie success rate prediction is not only useful for the producer or distributor but also to the movie lovers who don’t want to spent their time and money on movies which may not be interesting to them.

(Mahajan et al., 1984) dealt with the acceptance and distribution of new products and found that negative message passing plays a major role in success or failure of a product. The proposed approach will help to know which movie will succeed and which movie will fail before the release of the movie. In our approach we have made a detailed study of various classifier algorithms and their performance was evaluated in terms of their precision, recall and accuracy.

The paper is organized as follows: Section 2 elaborates the research work carried out related to this work which helped to enhance the work. Section 3 provides detailed design of the proposed model. Section 4 discusses about the experimental results and performance measures used for evaluating the proposed methodologies. Section 5 concludes the work done.

II. RELATED WORKS

(Vr et al., 2017) predicted movie success based on the data extracted from distinct sources IMDb, Rotten Tomatoes and Wikipedia. For carrying out the regression process, all nominal attributes are transformed to numerical attributes. Support Vector Machine (SVM) is often applied on classification and regression problems. In this work, SVM regression find a function from all the training data sets which has maximum deviation from the targets y. The process uses a linear kernel function to map the data into a high dimensional feature space and the linear regression is performed on that space. As Training data is minimum when compared to the number of features, the process used a linear kernel function. The limitation in this work is the usage of SVM for regression but no other efficient methods.

(Gothwal et al., 2018) predicted movie success by using the k-NN algorithm. Initially input consists of the k closest neighbors in the feature space. An item is classified by a majority vote of its neighbors, with the item being allocated to the class most normal among its k closest neighbors. Object is assigned to the class of that single nearest neighbor if k = 1. In k-NN regression modeling, the output is the average of the values of its k nearest neighbors in the feature space. The constraints in this work are k-NN has numerous exceptions and exactness is less when contrasted with other AI calculations, for example, Adaboost.

(Pramod and A, 2017) dynamically categorized the movies into the fuzzy sets without manual based checks so that the ability and efficiency of the engine got improved. The result of this method contains Original IMDB score, algorithmically determined IMDB score, categorization of the motion picture's prosperity, Percentile inside that class, Accuracy scope and Fuzzy Success. Despite the reality that their performance has shown a high degree of predictability, their calculation has downsides of terrible time unpredictability, as the fundamental information retrieval sets aside a lengthy attempt to create a training dataset for even a few tuples of data.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author
Balaganesh N*, Department of Computer Science and Engineering Mepco Schlenk Engineering College, Sivakasi, TamilNadu, India.
(email: balaganesh@mepcoeng.ac.in)
Bhuvaneswari M S, Department of Computer Science and Engineering Mepco Schlenk Engineering College, Sivakasi, TamilNadu, India.
(email: bhuvaneswari@mepcoeng.ac.in)

Verification and validation of a knowledge-based system needs comprehensive hardware testing that is virtually impossible to implement and operate.

A neural network based approach is proposed by Sharda and Delen (2006) for predicting the movie success rate. Eliashberg et al., (2000) proposed an strategy that would be helpful for producers and distributors to predict the success of the film before the original release.

Neelamegham and Chintagunta(1999) proposed Bayesian approach for forecasting the performance of the new movie. Asur and Huberman (2010) proposed a method for predicting the success of the movie grounded on the sentiments mined from the tweets. Latif and Afzal(2016) worked on various machine learning techniques such as simple logistic and logistic regression for predicting the success of movie using IMDB dataset. In our proposed model we have used Adaboost (Dietterich, 1997), SVM and kNN for movie success prediction(Hmeidi et al., 2008)

III. SYSTEM DESIGN

The detailed design of the entire process involved in the movie success rate prediction is given in Figure 1.

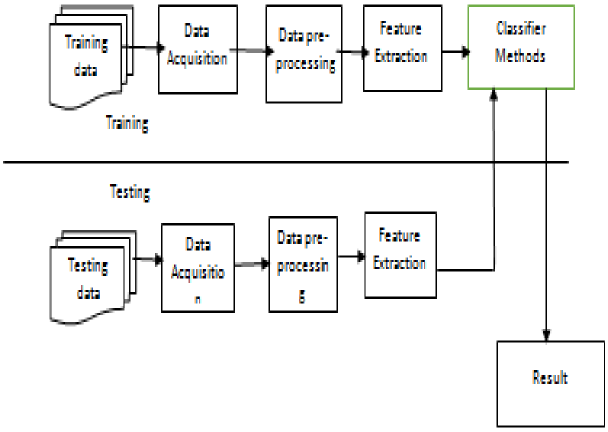


Figure 1: System Design

The modules present in the proposed model are as follows:

- Data Acquisition
- Data preprocessing
- Feature Extraction
- Classification

3.1 Data Acquisition

Data acquisition concerns information collection, although techniques may vary based on the sector, the emphasis stays the same on ensuring precision. Any data accumulation's vital goal is to capture quality data or evidence that efficiently implies a wealthy inquiry of data that can prompt sound and unquestionable answers to the issues that have been submitted. Precise accumulation of data is essential to ensure that the examination is upright, paying little regard to the field of research or data inclination (quantitative or subjective). The selection of suitable collection of data and tools, which could occur, adjusted or completely new, and with clearly defined instructions for their appropriate use, reduces the chances of errors occurring in the midst of accumulation.

Misshaped discoveries are frequently the after effect of information gathering, for example, deceiving inquiries on polls, accidentally discarding the accumulation of some supporting information, and other unexpected mistakes. This would prompt a skewed end that might be pointless.

Data acquisition enables one to address substantial inquiries and evaluate results. It is a research segment in all study fields, including physical, sociology, humanities, and business. While strategies shift through control, the emphasis on ensuring accurate and legitimate accumulation continues as before.

3.2 Data Preprocessing

Data preprocessing is a technique of data mining that involves converting raw data into a comprehensible form. The information used in the real world is incomplete, inconsistent and may contain some mistakes at times. Preprocessing information must therefore be performed to resolve the above-mentioned problems.

Data preprocessing techniques involves:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Data Discretization

3.2.1 Data Cleaning

Data cleaning is the way to recognize corrupted or incorrect records from a record collection, table or database and to differentiate fragmented, erroneous, unimportant parts of information and to replace, modify or delete dirty or coarse data.

3.2.2 Data Integration

Data Integration is the method of combining information from different sources into a meaningful form that the user needs. The most important part of data integration is to build an enterprise data warehouse. It will assist to analyze the information stored in the data warehouse and to draw significant findings and observations.

3.2.3 Data Transformation

The method of turning information into another type in one form is data transformation, which is simple to manage and analyze. In order to make it compatible with other data or to aggregate data information, data transformation is done. Since data often resides in various locations and formats, data transformation is necessary to ensure that data from one application or database is intelligible to other applications and databases.

3.2.4 Data Reduction

In the data warehouse, a vast quantity of information is stored.

Therefore, any information operation is hard to conduct. The data stored in the data warehouse must therefore be reduced.

Data reduction can be accomplished in cases where vast of information are needed to converted into small amounts of data but still critical information must be maintained.

3.2.5 Data Discretization

Data discretization is the process of transforming continuous value into a specified set of values/interval. Data correlation, clustering are some of the common methods of discretization. It will help to reduce the large amount of data within a particular set of values.

3.3 Feature Extraction

The output from data pre-processing is fed as input to Feature Extraction module. Once pre-processing of information is performed, extraction methods of characteristics are implemented to obtain characteristics that are helpful in classification. As features define the behavior of a data, they indicate their position in terms of storage, classification effectiveness and, clearly, time consumption. Feature Extraction can be defined as "extracting from the raw data information that is most suitable for classification purposes, while minimizing the within class pattern variability and enhancing the between class pattern variability". The main objective of feature extraction is to extract a set of features that maximizes the recognition rate with the least amount of elements and generates similar features for a variety of instances of the same type.

Various features are available, of those only certain numerical data values are extracted for predicting the result. The features that are considered in predicting the success of movie are runtime, rating, votes, revenue and genres.

3.4 Classification

Grouping is the way to predict the data set class. Approximating a mapping feature f is assigned from info variables x to discrete output factors y. The two types of learners in classification are lazy learners and eager learners.

Simply store the training data by lazy learners and keep up until a test information appears. When it does, grouping is resulted depending on the data stored preparing the most associated information. In contrast to eager learners, they have less time to prepare and more time to predict. Before receiving data for grouping, Eager Learners create a characterization model that depends on the data provided for preparation. Due to the growth of the model, eager learners set aside a lengthy preparation effort and less time to predict.

Different classifiers work by contrasting perceptions with past perceptions by methods for a likeness or separation work. The 3 classification methods adopted are

- Adaboost Classification
- Support Vector Machine (SVM)
- K Nearest Neighbour (KNN)

3.4.1 Adaboost Classification

Boosting is a method of general troupe making a strong classifier from different frail classifiers. This is completed by structuring a model from the training information, creating a second model at that stage that attempts to tackle the main model's errors. Models are included until the preparation set is superbly anticipated or as many models as possible are included. Using the weighted examples, a weak classifier is set up on the preparation test information. Just binary classification problems are reinforced and predictions are made by checking the weighted normal of the weak classifiers. For additional fresh input information, the expected value of each soft learner is either $^{+}$ 1.0 or -1.0.

For the trained model misclassification rate is calculated using Equation 1

$$error = \frac{correct - N}{N} \qquad (1)$$

Where, correct is the number of training instances that the model properly predicts, error is the misclassification rate, and N is the total training instances. Weighting for any prediction that the model makes can be calculated by using stage. Stage is calculated using Equation 2.

$$stage = \ln {1 - error}/_{error} \qquad (2)$$

The ensemble model's forecast is taken as the sum of weighted projections. If the sum is positive, the first class is predicted otherwise the second class is predicted to be negative.

3.4.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) is a supervised machine learning algorithm. It is used to fix issues of classification as well as regression. The algorithm produces an ideal hyperplane that categorizes fresh instances when a marked training data is provided as input. Divide the plane into two sections where there are two sides in each class.

SVM algorithm has the following tuning parameters:

- Kernel
- Regularization
- Gamma
- Margin

For linear kernel, the new input is predicted using Equation 3.

$$f(i) = C(0) + \text{sum}\ (c_i{*}\ (i, s_i)) \qquad (3)$$

i = input

s_i = support vector

C(0) and c_i = coefficients estimated from the training data by the algorithm

The Regularization parameter helps the SVM in avoiding the misclassification of each training sample. The size of the hyperplane margin is selected based on the value of the Regularization parameter. The Gamma parameter is used to define the influence of a single training sample. Low value for gamma parameter denotes low level of influence and high value for gamma parameter refers to high level of influence. The points near to the separation line are considered when the gamma value is high. Far away points are also considered when the value for gamma is low. A margin is a dividing line to the nearest class points. A excellent margin is acquired when there is a large amount of separation.

3.4.3 K Nearest Neighbor (K-NN)

K-NN algorithm is used for classification and regression. Predominantly it is used for classification. In K-NN classifier, the input contains the k nearest data in the feature space and the output is the class membership. In K-NN, k denotes the number of nearest neighbor. k is an odd number if there are two classes. In order to find the closest similar points, we have to use distance measures such as Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance to find the distance between points. KNN has the following basic steps:

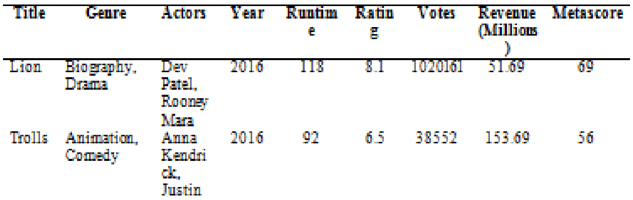
- Distance calculation
- Finding closest neighbors
- Vote for labels

There is no ideal number of neighbors in K-NN that matches any type of dataset. Each dataset has its own needs. If the number of neighbors are low, the noise will have a greater impact on the result and a large number will be computationally costly.

V. RESULTS AND DISCUSSION

The dataset taken into consideration is IMDb dataset. The sample of the dataset is given in Table 1.

Table 1: IMDb Dataset



Title	Genre	Actors	Year	Runtime	Rating	Votes	Revenue (Millions)	Metascore
Lion	Biography, Drama	Dev Patel, Rooney Mara	2016	118	8.1	1020161	51.69	69
Trolls	Animation, Comedy	Anna Kendrick, Justin	2016	92	6.5	38552	153.69	56

Data Preprocessing

The raw IMDB movie dataset is taken as input for cleaning process. Preprocessing is done to remove the inconsistencies in data and to transform the raw data into an understandable form. Initially there were 850 rows in the dataset in which some of the data have null values. Those rows which has null values are removed. Thus the elimination of the unavailable movie data is done. Initially in IMDB there was no revenue field for some movies, so then it was taken from websites like rotten tomato, box office mojo for the revenue attribute. Finally there were 839 rows in the dataset. The collected dataset is pre-processed by removing the names of directors, actor, and actress since all the names cannot be converted into a categorical value there is no point in using those names to classify the label.

The movies in the dataset have various genres so it is difficult to classify by treating them as such. So the genre was converted into numerical values. It had about 20 genres. If a movie falls under a particular genre then it is assigned a value of 1 otherwise the movie is assigned with a value of 0.

Classifier Model

The pre-processed information is divided into 1:3 ratio test and training data. Training dataset and training class label are assigned with 75% of data testing data and testing class label are assigned with remaining 25%of data. All the numerical values are normalized. Adaboost, SVM, KNN algorithms are applied to training set to identify the label, then the model built is used to predict test information class labels.

The training dataset is normalized and represented in array of vectors.

Adaboost

The weak classifier that is used in Adaboost classifier is decision tree classifier with maximum depth 2 in conjunction with support vector machine classifier. Adaboost algorithm has a very high accuracy of 99.5%.

Confusion Matrix :

$$\begin{pmatrix} 172 & 1 \\ 0 & 37 \end{pmatrix}$$

Accuracy, Precision and Recall obtained using Adaboost is shown in Figure 2.

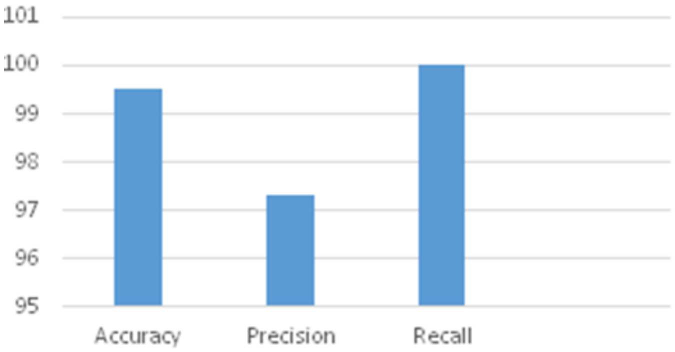


Figure 2- Line graph showing precision, recall and accuracy of Adaboost algorithm

The ROC (Receiver operating characteristic) curve is a representation of the true positive as opposed to the false positive for possible distinct points in a test. It shows that the specificity and sensitivity compromise. The specificity reduces when the sensitivity rises. The test's accuracy can be seen if the curve is nearer to the graph's top and left, closeness gives better results. If the curve goes close to the angle of 45 degrees, the findings would not be accurate. If the value of the ROC curve is above 0.9, the results are outstanding, 0.8-0.9 is great, 0.7-0.8 is fair and 0.6 is poor.

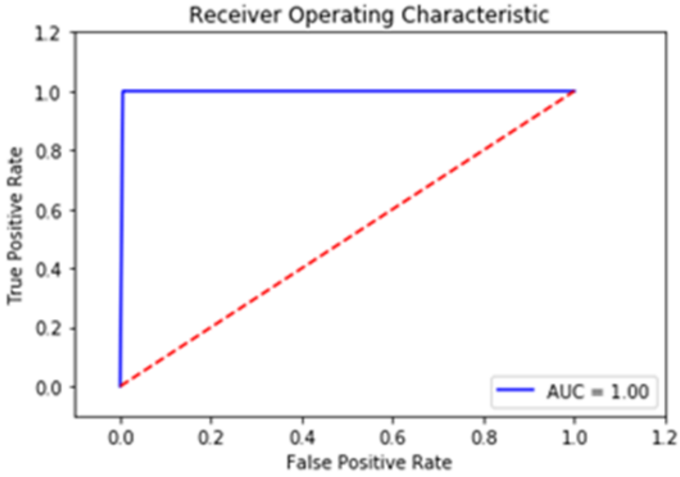


Figure 3– ROC curve for Adaboost algorithm

SVM

Each data item in this algorithm is plotted as a point in n-dimensional space with the value of each function being a coordinate value.

Confusion Matrix:

$$\begin{pmatrix} 166 & 7 \\ 10 & 27 \end{pmatrix}$$

Accuracy, Precision and Recall obtained using the algorithm is shown in Figure 4.

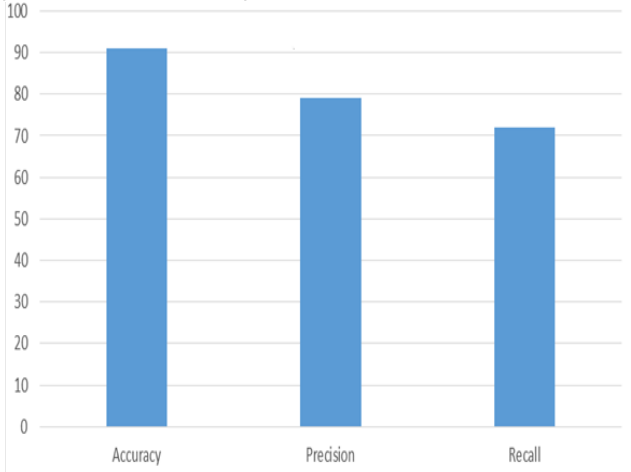


Figure 4 - Line graph showing precision, recall and accuracy of SVM algorithm

Figure 5 shows the outcome of true positive vs. false positive. This ROC curve gives an AUC of 0.84 which is better than the result of KNN classifier.

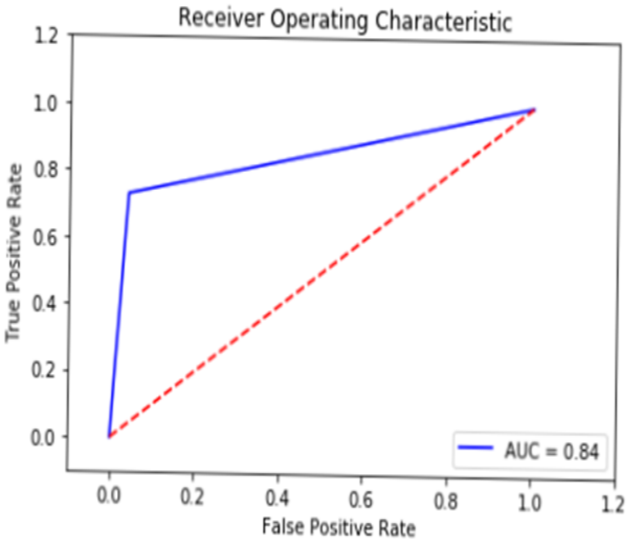


Figure 5 - ROC curve for SVM

kNN

The K-Nearest Neighbor classifier is used with k = 5 in which k is number of nearest neighbor.

Confusion Matrix :

$$\begin{pmatrix} 169 & 4 \\ 18 & 19 \end{pmatrix}$$

Accuracy, Precision and Recall obtained using kNN is shown in Figure 6

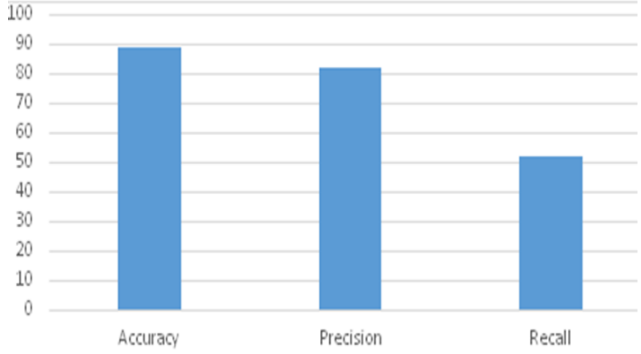


Figure 6 - Line graph showing precision, recall and accuracy of kNN algorithm

Figure 7 shows the result of true positive rate against false positive rate. ROC curve gives an AUC of 0.75.

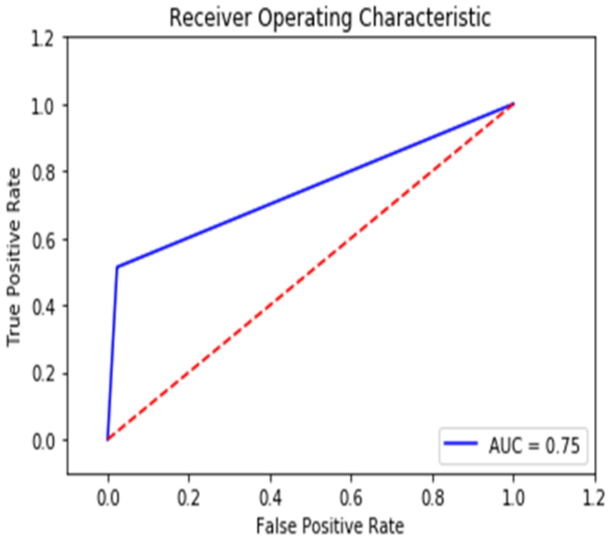


Figure 7 - ROC curve for kNN

Based on the performance measures Adaboost algorithm found to be the best choice for predicting the success rate of the movie.

VI. CONCLUSION

Success of a movie not only relies on attributes related to movie. There are various factors which determines movie success. The story of the movie, competitor movie which is going to be released along with the movie plays a vital part in determining the movie success. The budget of the movie, the director and actors/actress of the movie will also play a part. The main factor in improving model efficiency and achieving better results is a bigger training set. In this study three models were developed to predict the success rate of the movie, among which Adaboost classifier algorithm gives a better accuracy than SVM and KNN classifier algorithms. The proposed model will help the business people involved in cinema industry invest their valuable money on a movie. In future, additional attributes such as age of viewers and voters, current trends, analysis of news, movie plot analysis and analysis of social networks can also be added to improve the accuracy.

REFERENCES

1. Asur, S., Huberman, B.A., 2010. Predicting the Future With Social Media. https://doi.org/10.1109/WI-IAT.2010.63
2. Dietterich, T.G., 1997. Machine-Learning Research 18, 97-136.
3. Eliashberg, J., Jonker, J., Sawhney, M.S., Wierenga, B., Eliashberg, J., Mohanbir, J.J., Berend, S.S., 2000. MOVIEMOD?: An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures MOVIEMOD?: An Implementable Decision- Support System for Prerelease Market Evaluation of Motion Pictures.
4. Gothwal, K., Sankhe, D., Waghela, N., Sharma, M., Yadav, R., 2018. Movie Success Prediction 6, 66-69.
5. Hmeidi, I., Hawashin, B., El-qawasmeh, E., 2008. Performance of KNN and SVM classifiers on full word Arabic articles 22, 106-111. https://doi.org/10.1016/j.aei.2007.12.001
6. Latif, M.H., Afzal, H., 2016. Prediction of Movies popularity Using Machine Learning Techniques.
7. Mahajan, V., Muller, E., Kerin, R.A., 1984. Introduction Strategy for New Products with Positive and Negative Word-of-Mouth.
8. Neelamegham, R., Chintagunta, P., 1999. A Bayesian Model to Forecast New Product Performance in Domestic and International Markets 18, 115-136.
9. Pramod, S., A, G.M., 2017. Prediction of Movie Success for Real World Movie Data Sets 3, 455-461.
10. Sharda, R., Delen, D., 2006. Predicting box-office success of motion pictures with neural networks 30, 243-254. https://doi.org/10.1016/j.eswa.2005.07.018
11. Vr, N., Pranav, M., Pb, S.B., Lijiya, A., 2017. Predicting Movie Success Based on IMDb Data 3-6.