




# Popularity prediction of movies: from statistical modeling to machine learning techniques

Syed Muhammad Raza Abidi<sup>1</sup>  · Yonglin Xu<sup>1</sup> · Jianyue Ni<sup>1</sup> · Xiangmeng Wang<sup>1</sup> · Wu Zhang<sup>1,2</sup>

Received: 24 January 2019 / Revised: 9 October 2019 / Accepted: 27 November 2019

Published online: 06 January 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

## Abstract

Film industries all over the world are producing several hundred movies rapidly and grabbing the attraction of people of all ages. Every movie producer is of keen interest in knowing which movies are either likely to hit or flop in the box office. So, the early prediction of the popularity of a movie is of the utmost importance to the film industry. In this study, we examine factors inside the hidden patterns which become movie popular. In past studies, machine learning techniques were implemented on blog articles, social networking, and social media to predict the success of a movie. Their works focused on which algorithms are better at predicting the success of a movie but less focused on data and attributes related to an ongoing movie and in various directions. In this paper, we inspect this perspective that might be related to the prediction of the results. Data collected from the publicly available Internet Movie Database (IMDb). We implemented five machine learning algorithms, i.e., Generalized Linear Model (GLM), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), and Gradient Boosted Tree (GBT) using Root Mean Squared Error (RMSE) as a performance metric and got the accuracy performances of GLM: 47.9%, DL: 51.1%, DT: 54.5%, RF: 50.0%, and GBT: 49.5%, respectively. We found that GLM is the high achieving accuracy regression classifier due to the lower value of RMSE, which is considered to be better.

**Keywords** Movie popularity · Machine learning · Movie success · Regression · IMDb · Supervised learning

---

✉ Syed Muhammad Raza Abidi  
razaabdi@live.com; abidi@i.shu.edu.cn

✉ Wu Zhang  
wzhang@shu.edu.cn

# 1 Introduction

The Internet is the best source of information nowadays and, remarkably, every field massively uploading data over the internet and enormously becoming fast and efficient. The movie industry is also producing extensive data related to *stars, directors, studios, critics score, ratings*, and much more over the web, and it facilitates researchers to mine the data, trace, and identify the hidden patterns inside this big data related to movies [23]. Producing a successful movie is not an easy task for moviemakers. Focus only on some of the factors, such as genre, casting, and starring of a movie, could not enough due to the varied liking of the audiences. Other factors (such as director, famous actor/actresses, genre, and the cost) are considered as conventional factors, and also some non-conventional factors (such as movie trailer views on YouTube, likes on Facebook, and fan following on Twitter) make a movie successful [29].

Motion Picture Association of America (MPAA) declares that the growth and enlargement of the movie industry is a global phenomenon, and due to the influence of the movie industry on the economy of the country, many studies have conducted by scholars using machine learning techniques to predict the box-office success. Due to the prediction perspective, these researches have an essential effect on the movie industry [24]. The movie industry releases thousands of movies each year. As per the findings, in the United States, the movie industry produces profit up to 10 billion dollars, and almost every movie costs about 100 million, but despite their cost and production, still there exist some ambiguities and vagueness that either the movie will do business or not? [19].

As per the business perspective, the movie industry is one of the highest revenue-generating businesses. Of course, the success of a single movie can earn millions of dollars of profit for a studio, and the moviemakers are excitedly interested in making revenues from the movies through early predictions as a movie gets popular in the public results gross revenues from the community as well [48]. Most of the people have their hobbies to watch movies, and they are crazy about it. The movie is an excessive source of enjoyment, and people love to watch in the theater as well. Due to the liking of the divergent audience, the movie industry produces thousands of movies of diverse genres (such as *Action & Adventure, Mystery & Suspense, Science Fiction & Fantasy, Comedy, and Documentary*) every year [23].

Hollywood is the land of intuition, as the bulk of movies of varied interests and topics released every year in the United States. The situation is still unclear and uncertain for the studio that a movie will be successful or not, and this leads to the thought of prediction of movie success before its release [19, 47]. Cizmeci and Oguducu reveals that exposing the significant factors before releasing a movie could aid to box-office success. The producer and the other film making personnel could make proper decisions to make a movie hit. For example, if the movie becomes successful, then more audiences could watch it in various theaters, and the revenue will surely increase [10].

Several studies have conducted to predict the popularity of the movie. Mostly include user rating scores, while others used social media to predict the movie, e.g., Facebook, Twitter, and YouTube. Though on the other hand, limited work had conducted by considering movie features, such as *dates, Oscar-winning stars, director, studio, and runtime* for prediction of the movie [23]. In the study of Tang and colleagues, they evaluated movies with the same genre from IMDb and DouBan (Chinese social networking service). Initial findings could not produce solid evidence in support of the influence of foreign language on the popularity of the movie due to the limitation of the data. 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 [51].

According to Wang and Zhang, movie genres contribute a vital role in the popularity of the movie because the movie industry firmly makes decisions on what type of movie customers of different ethnic groups liked, rate, and favored the movie. Producing revenue generated movie is the eventual goal of every movie industry, and they depend on various market segments and customers' likings [53]. As far as the prediction of the most wanted and likely movies is concerned, Netflix's algorithm is the best example of the supremacy of data analytics/mining in the movie industry as the algorithm accurately predicts which particular movie an individual customer wants to watch next [16].

On the other hand, the availability of sufficient data about the movies over the web prompts to inspect knowledge discovery/engineering, data mining, and also machine learning. Movie industry and film producers become unsure whether the movie will get fame or do business in the future or not. They always think about how to market the movie, which target market should be focused, when to release the movie and how to publicize it. It is the reason that predicting a movie before its release is of the utmost significance to the film industry [5, 19].

In the study of Lee and colleagues, they proposed such a model that can lessen and reduce the ambiguity in predicting the performance of a movie. They investigated, past research has conducted using machine learning techniques, presents an equally high level of prediction performance and accuracy. However, discovering important prominent features might be substantial to anticipate the achievement of the movie. The power of the prediction model presented in the study of Lee and colleagues becomes inadequate because they only used alteration of the algorithms rather than concentrate on the feature selection and extraction [24].

Quader and colleagues tested and compared seven machine learning algorithms to predict the box-office success of a movie. They predicted the profit value based on pre and post-release features taken into consideration. They also appraised and highlighted the other features, such as the number of audiences, the economic condition of a country, Law and order situation, total annual ticket sold, and Gross Domestic Product (GDP) of a country as well to make a better prediction of a movie's box office success [41].

The prime objective of the past researchers was to introduce new machine learning algorithms and test its performance only, although their efforts have donated to the substantial growth of the prediction accuracy. However, many factors and perceptions could be considered further to improve and enhance prediction accuracy. For instance; It is measurable to explore the hidden, unfamiliar, and unknown features. Other ones are feature selection, and feature extraction from the existing features as it is one of the most often commonly considered methods to advance the accuracy and interoperability of machine learning algorithms [24].

Motivated by these previous studies, our study aims to utilize and extract the relevant features from the IMDB data to further understand the popularity of a movie. We focus on the feature aspect approach to improving prediction accuracy in this study. Further, we investigated the use of statistical and machine learning modeling and compared them to identify which are the best fit for the regression problem. The models identify the different patterns in the data, where the patterns can be identified as reflecting essential factors of prediction. It can also quantify which predictor occurrence worthy of the movie's popularity prediction. Moreover, due to the massive movies' data, it is possible to gather more features by fine-tuning the input parameters and criterions.

The rest of this article arranged as follows. Section 2 describes the evaluation of past studies on predicting the success of a movie. In Section 3, dataset collection and preprocessing used in this paper given. In Section 4, we present the proposed statistical methods and modeling, and then in Section 5, we define the machine learning regression models. In Section 6, we evaluate

the results of the prediction model built, discuss the various performance metrics, and analyze the predictive performance. Finally, we leave the reader with concluding thoughts and future works in Section 7.

## 2 Related works

The initial works embraced the research steered by B.R. Litman [27] and explored the attributes and their effects on the performance of the box-office. Litman further examined, the attributes (*i.e., critics score, genre, cost of the production, suppliers, theater release date, and award taking history of actors*). The movie industry at the moment kept rising since Litman's study, and for the sake of success and popularity of the movie has been an exciting and emerging research area; therefore, enormous articles have published. Prag and Casavant [39] exhibited a keen curiosity in classifying the association amongst features such as *costs of marketing campaigns, MPAA ratings, sequels, and success of a movie*.

The authors of [7, 20, 38, 44, 54] identified two known problems; sparsity and cold-start always faced in a collaborating filtering approach. The sparsity issue happens when there are insufficient user ratings, and customer data are available. Performance and accuracy of the recommendation collected by survey results from limited users will be lesser than gained built on a large number of examples. The other problem is cold-start, and it arises when movies and new customers do not have adequate facts available in the recommendation system [25, 45].

Basuroy et al. [6] had examined how *critical reviews* affect a movie's success, set the actor's power and finances. The authors of [12, 14, 35] had observed the association between the *actor's star power and the performance of a movie*. Many researchers applied different machine learning methods to content-based filtering, *i.e., K-means, Neural Network (NNET), and Naïve Bayes (NB)*. For instance, the idea employed by the NB classifier aims to identify whether an item is desirable by inspecting attribute information [50, 57].

The prediction regarding success, popularity, and business of the movie relied on machine learning techniques, as these learning techniques have formed prediction models with modest stages of accuracy [13, 15, 47]. For example, [47] has implemented some machine learning algorithms such as *discriminant analysis, DT, logistic regression, and NNET* and inspected the performance to predict a movie's success. The predictors they have used to forecast the movie accuracy and performance are *actor's star value, genre, MPAA ratings, special effects, sequel, competition level, and the number of screens* on the initial day of the movie release. Statistically, nine output variables with the 36.9% of accuracy predicted by their most beautiful performing machine learning model.

Zhang et al. [59] has proposed the NNET multi-layer backpropagation that has a better quality enhanced neural network model offered by [47]. Their model acceptably categorized six output variables with 47.9% of accuracy. Eliashberg et al. [15] has predicted the movie's return on investment relied merely on its script information using a DT. Zhang and Skiena [58] used electronic media articles to predict the gross of movies. Asur and Huberman used data from social media and using sentiment analysis to predict the future of the movies concerning the box office revenue or business [3].

For instance, anticipating such movies that are highly predictable to succeed is one of the research type, [4] they had considered social media data, *i.e., Twitter*, to forecast a movie success and [33] had utilized blogs information to predict sales of a movie. Asad et al. [2] used IMDb data and from Box office mojo, and for predicting the movie, they implemented PART

and C4.5 concurrently with the correlation coefficient matrix as a measure. They formed two dataset pre, and post-release movies and an experimented with it.

Additionally, parallel work had obtained in [37], where they focused on and used social media YouTube and Twitter comments for a similar objective. Mestyán et al. [32] got articles from Wikipedia and presented the prediction of the popularity of a movie. The study demonstrates that by using these articles, one can get nearly future outputs. In this research, they used Pearson's correlation coefficient and linear regression. They took features of the movie, such as *genre*, *release date*, *stars*, and *director* from Metacritic, and also used financial data from box office (i.e., opening revenue, and budget) from the figures.

Babu [5] used movie data from two online website sources, i.e., IMDb and Rotten Tomatoes, and one from Wikipedia as well. Babu collected data and implemented machine learning algorithms, such as linear & logistic regression, and support vector machine (SVM). Du et al. [13] has predicted the box office achievement by estimating the performance of three machine learning algorithms, i.e., linear regression, SVM, and NNET, examining the feelings and opinions of the texts poled on Tencent Microblog.

Different studies have been carried out by many researchers for predicting the movie success, for example, *social media*, *blogs*, *electronic media*, *print media*, and *publications*, but still, there are shortcomings of researching features of a movie [23]. After getting succinct knowledge from the prior studies, some researchers have instigated to conduct the research that has a predictive nature. Mostly in past studies about the movie industry have had descriptive, illustrative nature, inspecting aspects, or features that disturb the box-office performances of movies [24].

Kim and colleagues applied lexicon-based sentiment classification and machine learning methods for predicting the success of a movie. They established a sentiment dictionary by using feature extraction and polarity assignment. Their findings showed a strong positive relationship between the sentiments of the audiences and box-office success. The relationship also significant and improved prediction accuracy by using a linear regression model [22].

Wang and Zhang [53] used the two approaches in their research, i.e., collaborative filtering and content-based filtering. In collaborative approach items of attention to a specific user grounded on the resemblance to prior rating history, and in the content-based filtering method or approach, the procedure is constructed on details of items and user likings to recommend items to customers further. The method relates to the user's likings with illustrations of the new items and also matches with item features.

## 2.1 Literature review

The related terminologies used in past research for predicting the success of a movie enlisted in Table 1.

## 3 Dataset collection and preprocessing

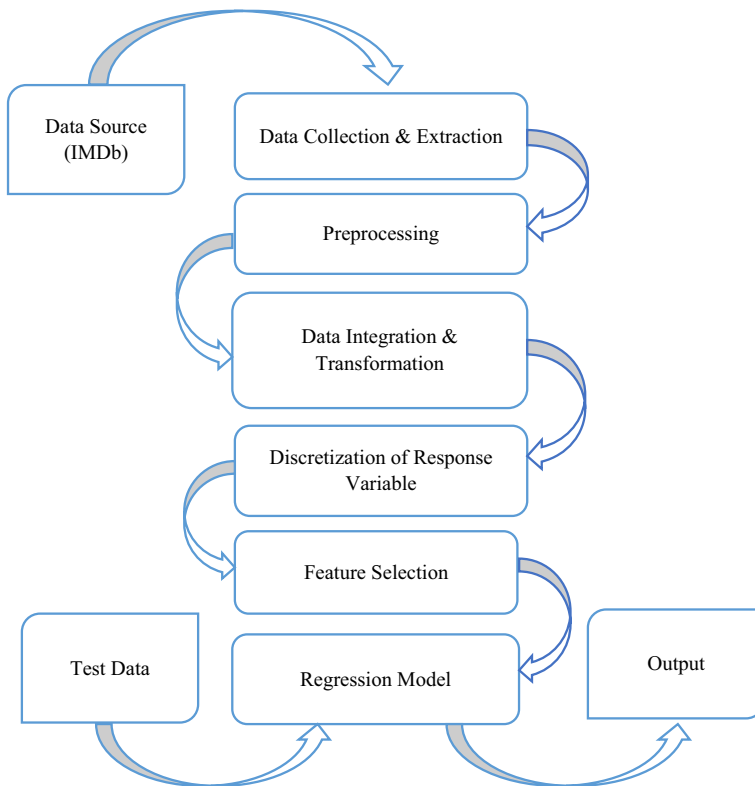
In this section, we explained the steps involving the data collection and preprocessing, which is an essential step before applying machine learning methods and techniques shown in Fig. 1.

### 3.1 Data collection & extraction

The dataset used in our study collected from the IMDb webpage, and it includes movies released from 1972 to 2014. For making more accurate predictions, we selected those movies that are listed

**Table 1** Summary of past techniques

Techniques used	Pros	Cons
Regression Analysis [6, 27, 32, 35, 37, 39]	A statistical technique used to estimate future situations offers a significant relationship between two or more related variables, through which one can predict the unknown values of one variable from the known values of another variable.	Lengthy and complex calculations, cause, and effect of the relationship among variables remain unaffected, cannot be used in a qualitative phenomenon.
Collaborative Filtering Algorithm [7, 20, 44, 45, 54]	Used by recommender systems, good for automatic predictions using previous preferences of users (likes or dislikes) data.	Not mainly succeed in automatically matching content to one's likings caused by the cold start problem.
Content-based Retrieval, and NB [25, 45, 50, 57]	In content-based filtering, items recommended by comparing the content with user interests. NB works on Bayes theorem, converged quickly than discriminative models, i.e., Logistic Regression, and require less training data.	If missing content or information is available, then do not perform precised recommendations. Make the dominant assumptions on the form of data distribution, data scarcity, and continuous features problems arise.
Logistic Regression, Discriminant Analysis, Classification and Regression Tree, SVM, NNET [13, 15, 47, 59]	Logistic Regression: Widely used technique, efficient to train, feature scaling not required. Discriminant Analysis: A statistical technique specifies models for predictors, stepwise & best subset analysis, helps for categorical regression analysis, and determine the accuracy among the groups. DT: Easy to implement and understand, suitable for nominal and numeric inputs, and to classify unknown outputs fast. SVM: It models non-linear decision boundaries, and robust against overfitting. NNET: It has fault tolerance, the capability to work with incomplete knowledge, distributed memory, and parallel processing ability.	Cannot solve non-linear problems, high dependence on the proper arrangement of a data, and explanatory variables not identified then gives poor performance. It cannot be used when subgroups are stronger; without strong classification, no predictors can be selected. If the number of observations or records is less, then the discrimination method cannot be used. The output variable must be categorical, and numerical input variables lead to complex trees. Memory intensive, trickier to tune picking the right kernel, and does not scale well for a large dataset. It has hardware dependence, unexplained behavior of the network, and difficulty of showing the problem to the network.
Regression and K-nearest Neighbor [58]	K-Nearest Neighbor: Simple to understand and easy to implement, Tags the new data entry-based learning from past data and continually evolves.	Declines performance when dataset grows, the curse of dimensionality, no capability of handling the missing values, and imbalanced data causes problems.
Classification using C4.5, and PART [2]	C4.5: Builds decision trees, deals with both nominal and numeric attributes, missing values, and pruning trees, and treats the training data example as a pair. PART: It is the algorithm of WEKA, builds partial C4.5 decision trees to generate a decision list, and makes the best leaf into a rule, and can also be used for feature selection.	May suffer overfitting, lots of classes makes it hard to learn or predict, and biased with those nominal/categorical variables which have more levels.
Linear Regression, Logistic Regression, and SVM [5]	Linear Regression: Linearity, predict the values of a response variable based upon the values of one or more explanatory variables.	Response variable must be continuous or at least close to continuous, only looks at the mean of a response variable, limited to linear relationships among variables, sensitive to outliers. If the number of features is more than the number of samples, the model starts to noise rather than the relationship between the variables.
Cinema Ensemble Model (CEM) [24]	Proposed for improvement of prediction accuracy, composed of seven machine learning algorithms, and focuses on the selection of features.	Weighted voting criterion did not consider, and other features or data may be used, i.e., Twitter.
Logistic Regression, Multilayer perceptron, J48, NB, and PART [23]	Multilayer perceptron: Multivariate non-linear regression models, uses backpropagation technique for training, can learn conditional probabilities, has multiple layers and non-linear activation, and able to generalize to new data. J48: C4.5 algorithm implemented in WEKA as a classifier called J48.	Choosing several hidden layers and nodes, i.e., How many are enough? Weight initialization, catastrophic forgetting when a trained network becomes further trained on new data, and generalization decrease when the number of connections approaches the number of training examples.
Logistic Regression, and Gaussian Kernel SVM [53]	Gaussian Kernel SVM: It is merely a weighted linear combination between a data point and support vector, a good classifier when the number of features is small.	Kernel models are sensitive to over-fitting, and not scale well for a large number of features or a large number of training samples.



**Fig. 1** General proposal

on Wikipedia list of years in film pages and are English movies released in the United States, and the rest excluded. We also removed movies which do not have any information about box office details.

The given data set consists of 651 randomly sampled movies produced and released. Data randomly sampled so; therefore, we can assume the generalizability of our conclusions. There is no random assignment used as it is observational data not experimental so, therefore, we cannot assume any causal relationship between the explanatory and response variable.

### 3.2 Data preprocessing

The data we attained from the available online database, i.e., IMDb, and need to be cleaned as the data are incredibly prone to noisy, and missing due to the massive size from a publicly accessible online source [17]. Initially, our data record was consisting of 651 rows with features related to movies as listed in section 3.3, Table 2. After cleaning of missing values by ignoring incomplete observations, such as, features with missing information represented by “N/A” or left blank wholly

**Table 2** Proposed movie popularity features

Class	Features
Nominal	genre, mpaa_rating, studio, best_pic_nom, director
Numerical	runtime, thr_rel_year, thr_rel_month, dvd_rel_month, dvd_rel_day, imdb_num_votes



deleted from the data set to avoid skewing the results. This initial round of cleaning provided 632 complete responses.

### 3.3 Data integration & transformation

The next step is, integration and transformation of the data into one database as data are coming from heterogeneous sources. Through this step, we can implement a statistical analysis and regression process more efficiently and quickly. Our dataset comprises both nominal and numeric attributes. For a regression process, we need all features to be numerical, and for this purpose, we used statistical programming language R (<https://cran.r-project.org/>) to accomplish this task. List of anticipated features shown in Table 2.

### 3.4 Discretization of the movie popularity

In this study, we define the prediction of the popularity of the movie as a regression problem. This approach applied in a few earlier studies, e.g., [5]. We discretize the dependent variable (i.e., `imdb_rating`) because it has continuous numeric values.

## 4 Research methodology

In this section, we describe the methodology behind experiments that performed.

### 4.1 Exploratory data analysis

#### 4.1.1 Selection of predictors

After setting a research question, we now turn to choose which variables to include in our model and eliminate or drop those variables which are not useful for our model. Table 3 reveals the reason for the rejection of other predictors.

So, after elimination, we are left with five nominal, and six numerical types of features in this study amongst 25 features shown in Table 2. Few features, we have selected including the ones that widely used in past studies. Besides, we have also nominated the features which corroborate statistically and are enough to predict the popularity of the movie successfully [24].

We have used R and RStudio (<https://www.rstudio.com/>) to convert the categorical/nominal features to some numeric values. It converts these values into binary features. A variable with

**Table 3** Excluded predictors

No.	Excluded predictors in the sample dataset	Reason
1	Movie URLs	Not significant for prediction
2	Title of the movie	As it introduces too much variability, as it does not have an exact string
3	Names of actor(s)/actress(es)	Too much variability for individual names to be significant enough
4	top200_box, best_actor_win, best_actress_win, best_dir_win, best_pic_win, thr_rel_day, dvd_rel_year, and title_type	For similar variability and distribution reasons



more than two possible values converted into n-binary features, where n represents the number of values. For instance, genre, one of the features in this study, has eleven possible values, including *Action & Adventure*, *Animation*, *Art House & International* [24].

The following sub-sections describe the nominal features included in this study.

**Genre** It is one of the most simple and frequently used variables in predicting a movie's success [47]. In this study, we used the eleven categories as follows: ACTION & ADVENTURE, ANIMATION, ART HOUSE & INTERNATIONAL, COMEDY, DOCUMENTARY, DRAMA, HORROR, MUSICAL & PERFORMING ARTS, MYSTERY & SUSPENSE, OTHER, and SCIENCE FICTION & FANTASY. The information on movie genres are available on the webpage of the IMDb.

**MPAA\_rating** Assigned by MPAA to the movie. A film rating system used in the United States. These ratings signify violence, sexual content, and language in a movie. There are six categories for each of the movies, mainly G, NC-17, PG, PG-13, R, and Unrated [23].

**Studio & director** The data about the studio & director of a movie, producing studio could be useful in modeling. There are too many values in the corresponding variables, e.g., WARNER BROTHERS PICTURES, twentieth CENTURY FOX, COLUMBIA PICTURES, DISNEY, HBO, PARAMOUNT STUDIOS, etcetera. Instead of using them directly, we are going to divide directors and studios into four ranks. A rank is a number from 0 to 3. If the average rating of movies for a studio or a director falls into the first quartile of the distribution of `imdb_rating`, we assign "Rank 0". "Rank 1" for the second quartile, and so forth. We need a function to determine the quartile of value for that. Since the distribution is not normal, we cannot use the theoretical method of determining quartiles. Instead, we are going to use the "ecdf" function in the R language.

**Best\_pic\_nom** This variable contains the two possible values "Yes," and "No." Table 4 shows the detail of the nominal predictors used in this study.

#### 4.1.2 Selection of (predicted) response variable

We are interested in learning what attributes make a movie popular – so, we have a few variables to choose from the list. Here are the details of the popularity related variables that are

**Table 4** Summary of the nominal predictors

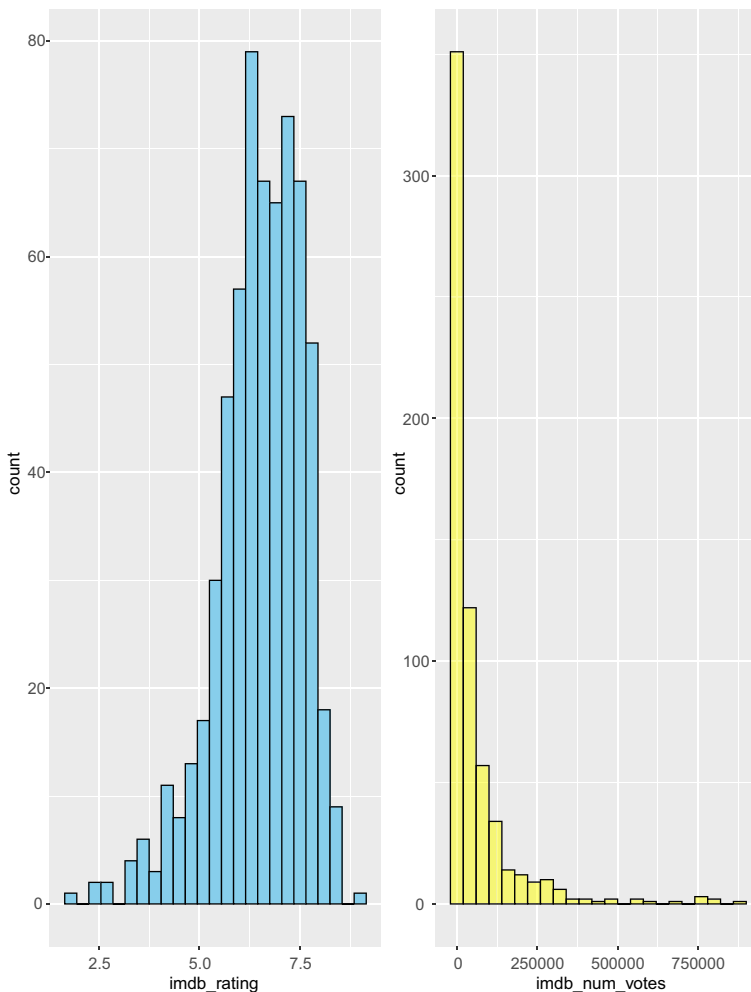
Name of predictor	No. of Instances	Values
genre	11	Action & Adventure, Animation, Art House & International, Comedy, Documentary, Drama, Horror, Musical & Performing Arts, Mystery & Suspense, Other, And Science Fiction & Fantasy
mpaa_rating	6	G, NC-17, PG, PG-13, R
studio_rank	0, 1, 2, 3	Warner Brothers Pictures, twentieth Century Fox, Columbia Pictures, Disney, HBO, Paramount Studios, etcetera.
director_rank	0, 1, 2, 3	Alex Smith, Brad Anderson, Cameron Crowe, David O. Russell, Edward Zwick, etcetera.
best_pic_nom	2	Yes, No

continuous numerical. For the regression model, we selected two features for the response variable:

1. `imdb_rating`: Rating on IMDb
2. `imdb_num_vots`: Number of Votes on IMDb

Both of these look-like legit measures of popularity, so, we will choose our response variable concerning their distribution only. We have used the "ggplot2" library in R to draw plots shown in Fig. 2.

Figure 2 proves that `imdb_rating` is closest to a normal distribution, which should contribute to the robustness of the model, so this shall be our response variable.



**Fig. 2** Statistical distribution of the response variable

## 4.2 Investigation and feature selection

There are many standards available for feature selection (such as *Backward Elimination*, *Forward Selection*, *Akaike Information Criterion (AIC)*, *Bayesian Information Criterion (BIC)*, *Deviance Information Criterion (DIC)*, *Bayes factor*, and *Mallow's Cp*). In this study, we used Backward Elimination using the adjusted  $R^2$  method to construct our model because it is a common way [52].

In this technique, we start with the full model and eliminate one variable at a time until the parsimonious model is reached [43]. In the end, features that are the redundant and weaker correlation with the response variable eliminated. Important steps involved in the Backward Elimination using the adjusted  $R^2$  method [5] shown in Fig. 3.

### 4.2.1 Explorations

After eliminating unwanted variables and choosing our response variable, we go ahead and get a feel for the data using some summaries and plots shown in Figs. 4, 5, 6, and 7. Figure 4, reveals the relationship is not very strong because there are some anomalies seen in the data of genre. The black dots show the outliers in the data, which accelerated the mean above the value of 6 in the `imdb_rating`.

The plot Fig. 5 below, assures good functional relationship as the runtime of the movie goes longer or more prolongs; the `imdb_rating` goes higher than the previous ratings.

Now, look at Fig. 6, which demonstrates the time factors, as we have two of these, i.e., `thr_rel_year`, and `thr_rel_month`. There appears to be some fan shape trend over the years, as variability grows slightly higher as years go along but no apparent trend within the months. It has shown in Fig. 7 separately, using a box plot which shows the outliers at or below 4 of the `imdb_rating`, and removing these outliers could illustrate at the mean value of `imdb_rating`. We saw some differences, but it does not seem like much to account for its significance without tests, which we have performed in section 4.2.3 Model diagnostics.

### 4.2.2 Statistical modeling

After a straightforward elimination of predictors, we ended up with the list of features and applied multiple linear regression model to achieve a model with a high adjusted  $R^2$  value. The

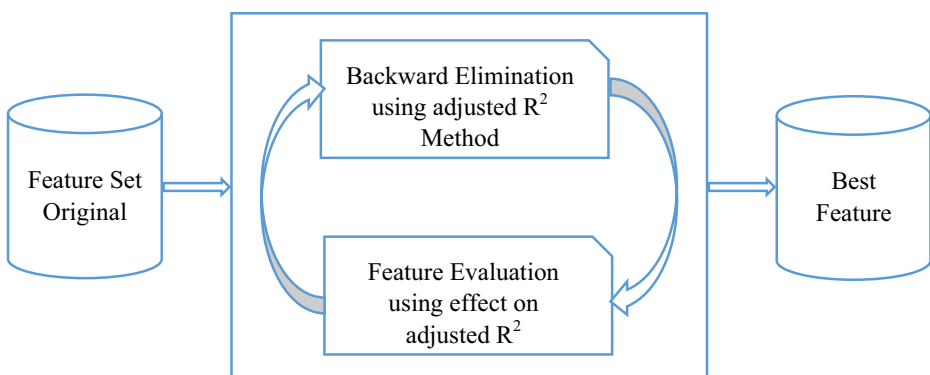
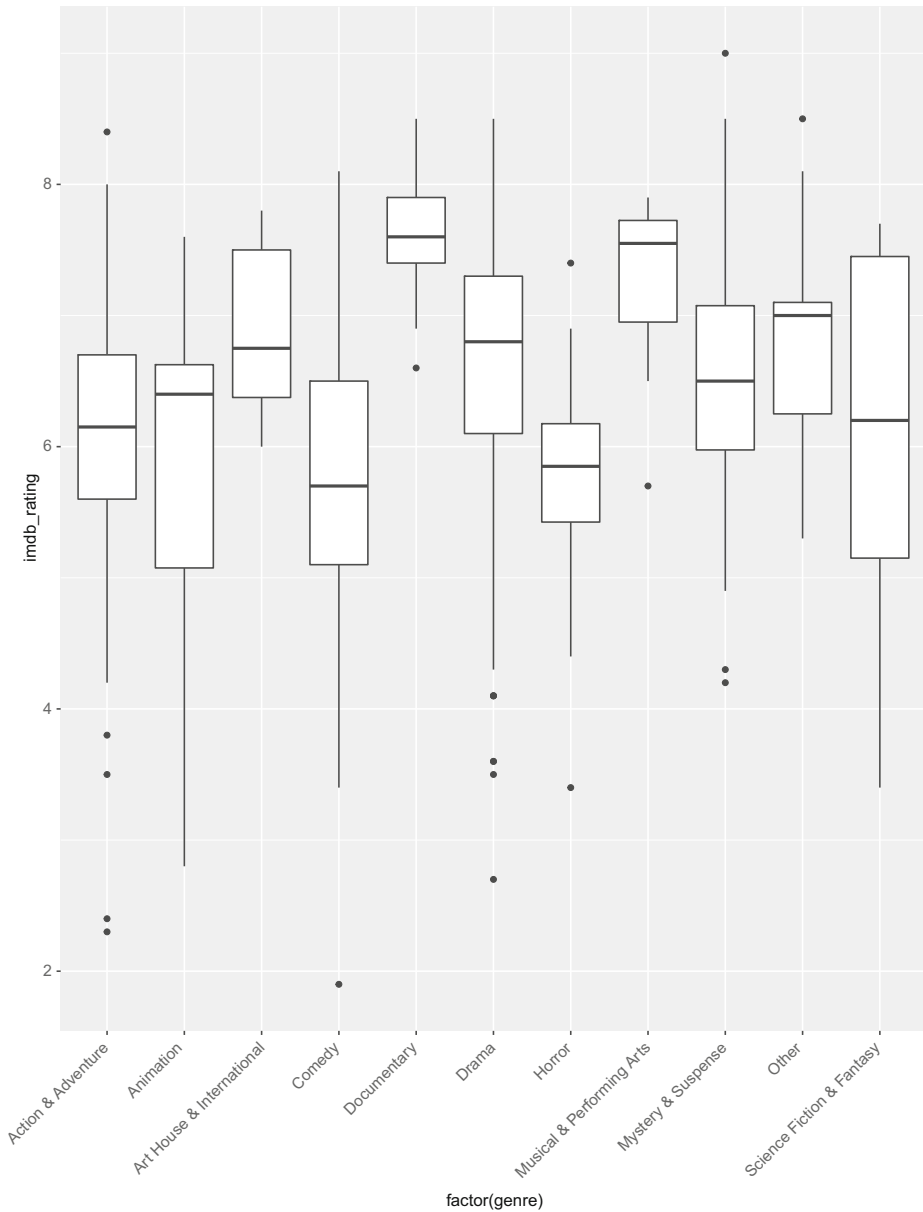


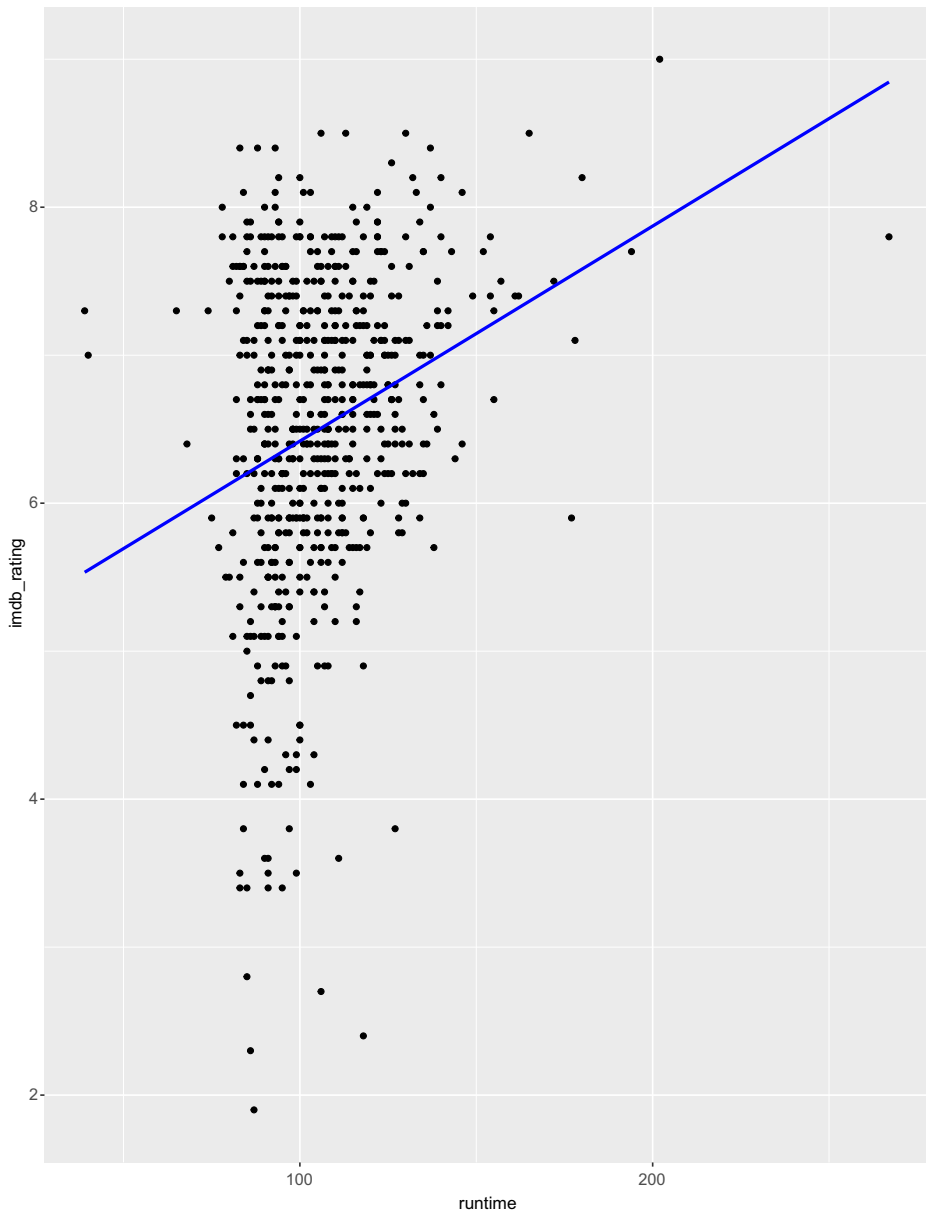
Fig. 3 Feature selection using a statistical technique



**Fig. 4** A boxplot shows the relationship between genre and imdb\_rating

technique starts with the set of all features. We iterate over the full features, at each iteration, it checks the adjusted  $R^2$  value, if it gets even a slightly greater change in the value, removes one of the collinear predictor variables remaining in the set.

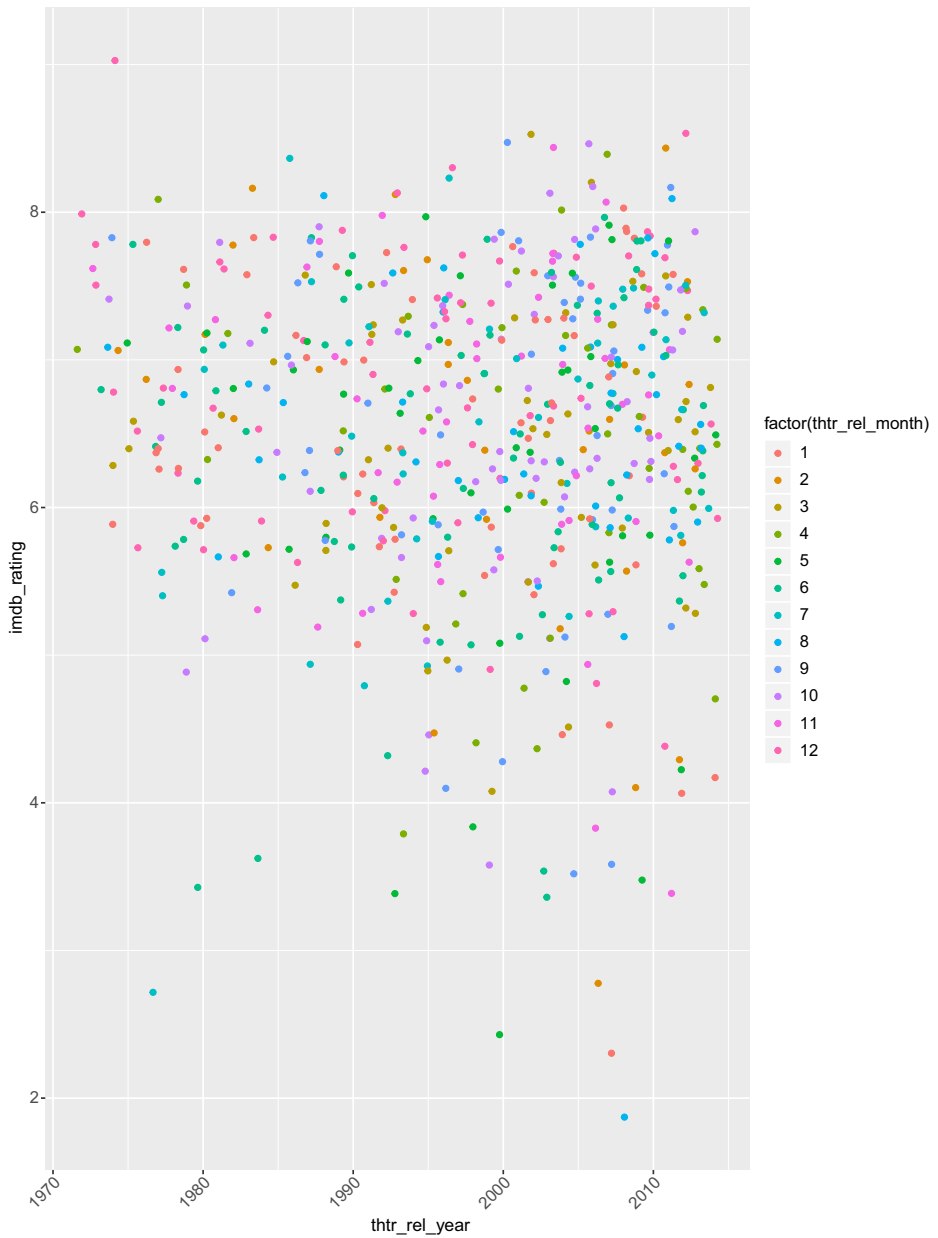
Finally, in the end, it gives a robust model with the assurance that all the predictors correlated with the response variable, and the redundant predictors eliminated. Table 5 reveals a summary of the final model.



**Fig. 5** A scatter plot showing the relationship between runtime and imdb\_rating

With regard to inference for the model, the *p*-value of the model's *F*-statistic indicates that the model as a whole is significant. It also noted that not all predictors have a significant *p*-value as the model was developed using the highest adjusted  $R^2$ .

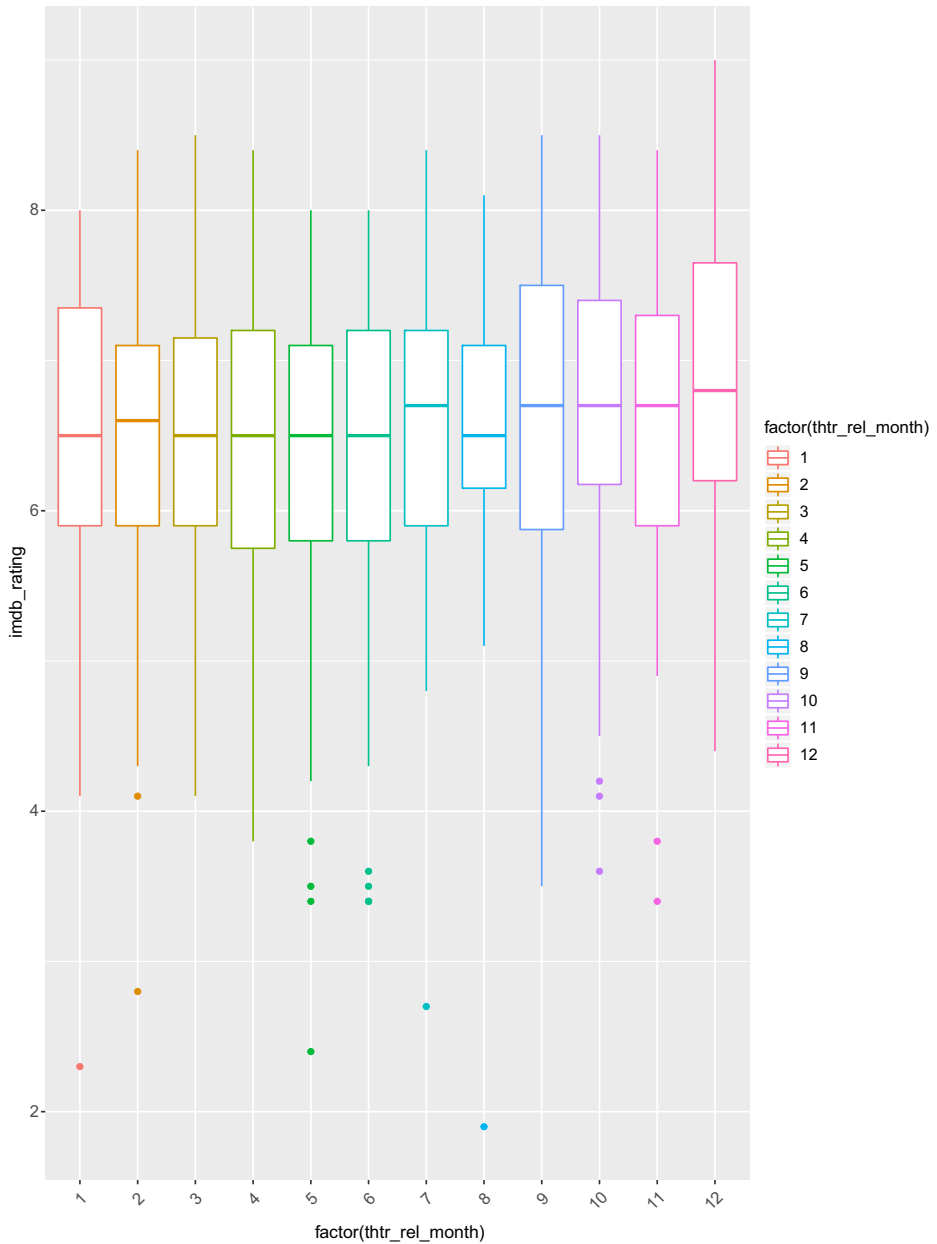
*Interpretation of the model coefficients* coefficient for director\_rank shows that for each unit increase in the value, the imdb\_rating is increased by approximately 6% with a very low *p* value, similarly for each unit increase in the value of studio\_rank, the imdb\_rating is increased



**Fig. 6** A scatter plot showing the relationship between `thr_rel_year` and `imdb_rating`

by approximately 1% with a very low  $p$  value as shown in Table 6. We might prefer to look at an ANOVA table too:

Here, we can see that *all independent variables are significant predictors* based on their  $p$ -values.



**Fig. 7** A box plot showing the relationship between `thtr_rel_month` and `imdb_rating`

#### 4.2.3 Model diagnostics

**Validity** In order for the multiple regression model to be valid, it is mandatory that the model should validate below four conditions:



**Table 5** Final multiple linear regression model interpretation

Residuals:					
Min	1Q	Median	Max	3Q	
-2.89508	-0.19224	0.05578	0.28837	1.83471	
Coefficients	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.099e+01	4.310e+00	4.869	1.43e-06	***
genreAnimation	-4.429e-02	2.104e-01	-0.211	0.833340	
genreArt House & International	2.854e-01	1.671e-01	1.708	0.088201	.
genreComedy	1.098e-01	8.686e-02	1.265	0.206517	
genreDocumentary	3.941e-01	1.222e-01	3.226	0.001325	**
genreDrama	2.735e-01	7.523e-02	3.636	0.000300	***
genreHorror	2.090e-01	1.307e-01	1.599	0.110405	
genreMusical & Performing Arts	3.805e-01	1.670e-01	2.278	0.023060	*
genreMystery & Suspense	3.204e-01	9.590e-02	3.341	0.000886	***
genreOther	2.965e-01	1.497e-01	1.980	0.048099	*
genreScience Fiction & Fantasy	-3.066e-01	1.925e-01	-1.593	0.111684	
Runtime	2.628e-03	1.265e-03	2.078	0.038171	*
mpaa_ratingNC-17	-2.128e-01	5.315e-01	-0.400	0.688949	
mpaa_ratingPG	-1.610e-01	1.459e-01	-1.103	0.270486	
mpaa_ratingPG-13	-2.660e-01	1.517e-01	-1.753	0.080074	.
mpaa_ratingR	-1.711e-01	1.457e-01	-1.174	0.240700	
mpaa_ratingUnrated	-2.697e-01	1.702e-01	-1.585	0.113528	
studio_rank	1.751e-01	2.933e-02	5.969	4.05e-09	***
thtr_rel_year	-8.162e-03	2.154e-03	-3.790	0.000166	***
thtr_rel_month	3.063e-03	6.118e-03	0.501	0.616846	
dvd_rel_month	1.593e-02	6.134e-03	2.596	0.009658	**
imdb_num_votes	1.135e-06	2.247e-07	5.052	5.80e-07	***
best_pic_nomyes	1.110e-01	1.215e-01	0.914	0.361251	
director_rank	6.820e-01	2.693e-02	25.321	< 2e-16	***
Residual standard error: 0.5098 on 608 degrees of freedom					
Multiple R-squared: 0.7835, Adjusted R-squared: 0.7753					
F-statistic: 95.64 on 23 and 608 DF, p value: < 2.2e-16					

Significance codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1, Annotation = \*\*\*: Significance level: 0.001,  $p$  value: [0, 0.001], \*\*: Significance level: 0.01,  $p$  value: (0.001, 0.01], \*: Significance level: 0.05,  $p$  value: (0.01, 0.05],.: Significance level: 0.1,  $p$  value: (0.05, 0.01], Blank space: Significance level: 1,  $p$  value: (0.1, 1]

1. There is a linear relationship between any numerical predictor variables (*runtime*, *thtr\_rel\_year*, *thtr\_rel\_month*, *dvd\_rel\_month*, *imdb\_num\_votes*) and the response variable (*imdb\_rating*).
2. The residuals are nearly normally distributed.
3. Residuals display constant variability and
4. The residuals are independent.

First, we will examine whether the binary variables included in the model are linearly related to the response variable or not? Figs. 8, 9, 10, and 11 demonstrates graphically and satisfies the above-stated conditions.

*Condition 1: Linear relationship between numerical ( $x$ ), and  $y$*

Figure 8, illustrates the *imdb\_rating* by examining the distribution of the residuals and observe whether the numerical variables included in the model are linearly related to the

**Table 6** Analysis of variance

Analysis of Variance Table Response: imdb\_rating

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)	
genre	10	166.887	16.689	64.2193	< 2.2e-16	***
runtime	1	34.646	34.646	133.3204	< 2.2e-16	***
mpaa_rating	5	15.985	3.197	12.3021	2.216e-11	***
studio_rank	1	114.456	114.456	440.4347	< 2.2e-16	***
thtr_rel_year	1	1.211	1.211	4.6585	0.0312908	*
thtr_rel_month	1	1.698	1.698	6.5337	0.0108271	*
dvd_rel_month	1	3.136	3.136	12.0670	0.0005497	***
imdb_num_votes	1	66.240	66.240	254.8973	< 2.2e-16	***
best_pic_nom	1	0.787	0.787	3.0300	0.0822429	.
director_rank	1	166.621	166.621	641.1689	< 2.2e-16	***
Residuals	608	158.002	0.260			

Significance codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1. Annotation = \*\*\*: Significance level: 0.001,  $p$  value: [0, 0.001], \*\*: Significance level: 0.01,  $p$  value: (0.001, 0.01], \*: Significance level: 0.05,  $p$  value: (0.01, 0.05],.: Significance level: 0.1,  $p$  value: (0.05, 0.01], Blank space: Significance level: 1,  $p$  value: (0.1, 1]

response variable. A residual is a difference between the observed value and the actual or theoretical value. Thus, Fig. 8 validates the condition 1.

*Condition 2: Nearly normal residuals with mean zero*

In Fig. 9, Histogram and Normal probability plots demonstrate the residuals are nearly normally distributed and satisfy the condition 2.

*Condition 3: Constant variability of residuals*

Figure 10, reveals the residuals' constant variability and allows for considering the entire model with all explanatory variables at once. It depicts in Fig. 10 and satisfies the condition.

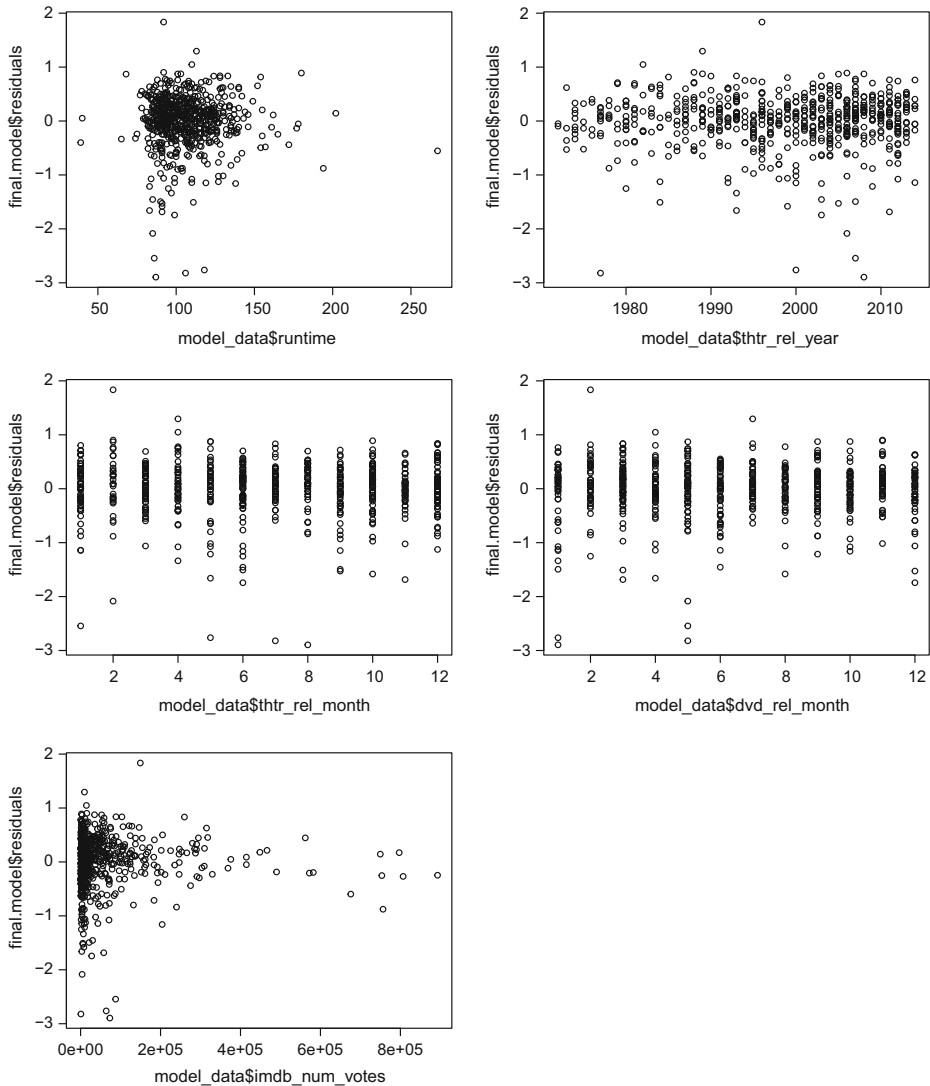
*Condition 4: Independent residuals if time series structure suspected*

Figure 11, confirms that the residuals are independent. As the plot shows the relationship of residuals among all the explanatory variables, and it seems near the mean with no fan shape presentation.

## 5 Construction of the machine learning regression model

The statistical analysis above satisfies that the data set we used is concrete and robust enough to implement the machine learning techniques on the given data. Although machine learning algorithms work on the principles of Statistics but performing Statistical tests and models is much better before applying machine learning techniques to the data.

In this study, we used a supervised learning technique as a response variable output is known. We used five machine learning methods to build candidate models for predicting the popularity of the movie and will compare the performance of different methods.

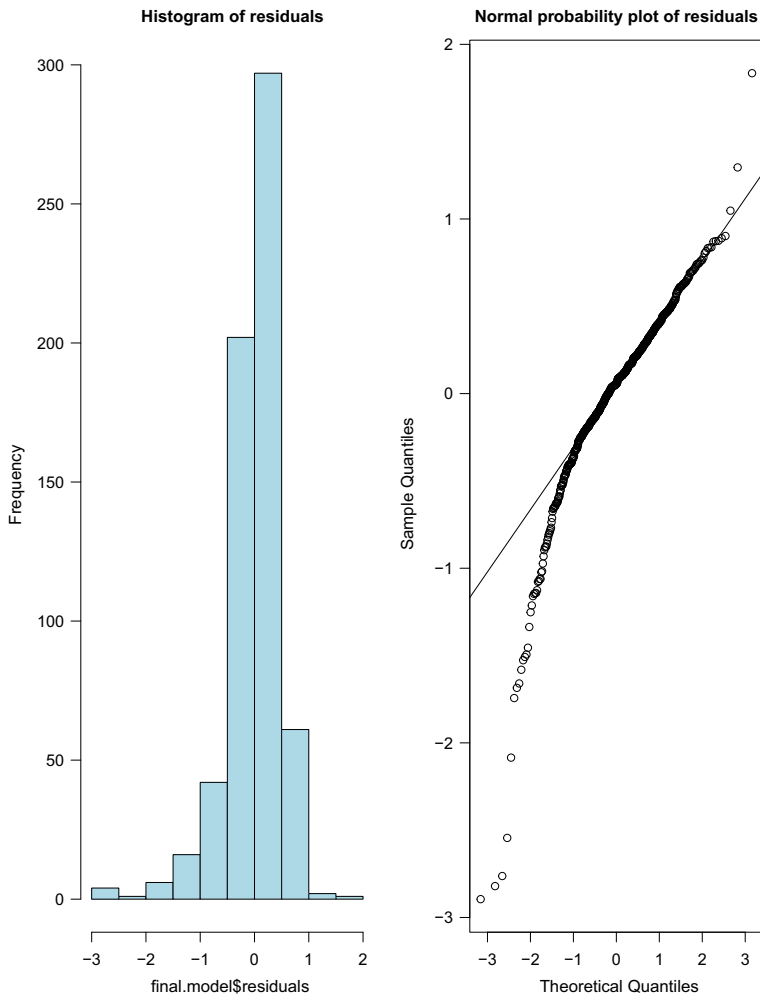


**Fig. 8** Linear relationship between numerical predictors and response variable

### 5.1 Generalized linear model (GLM)

It works and evaluates on maximum likelihood estimation (MLE), a well-known statistical principle. The primary objective of GLM is to minimize the difference between the actual and the forecasted value of the response variable, which is Gaussian distributed and called a residual [36].

GLMs are the augmentation of old-fashioned linear models, and these models use the series of commands by using the well-known MLE technique. These models are speedy and perform parallel computation even with a smaller number of predictors with non-zero constants [42].

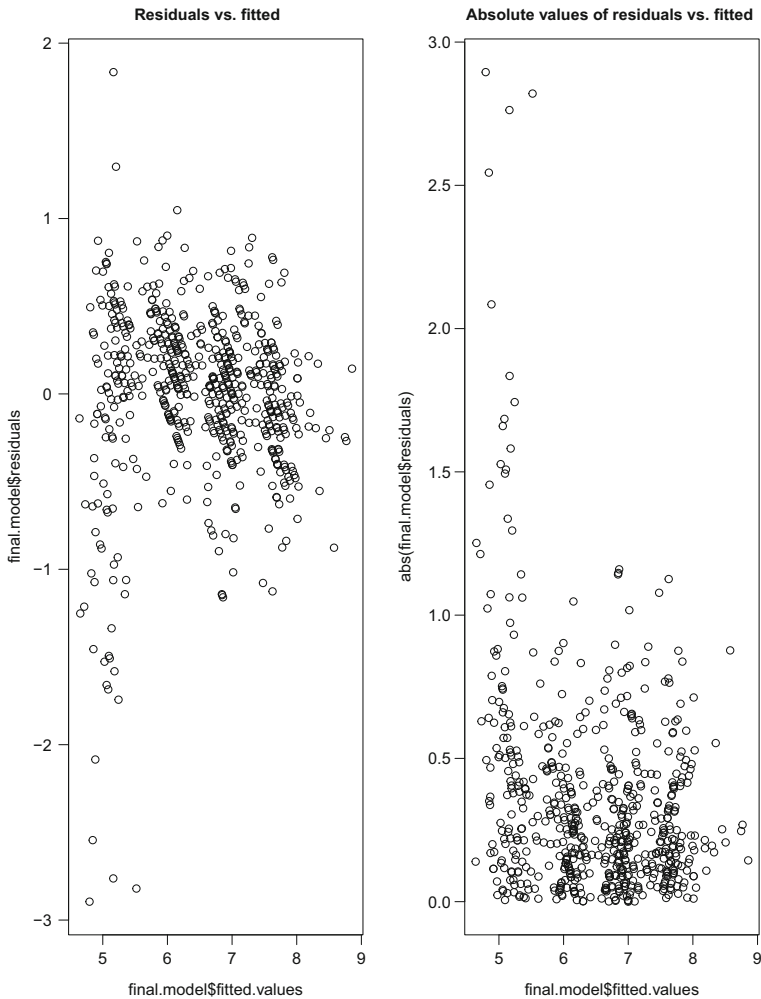


**Fig. 9** Distribution of residuals with mean zero

## 5.2 Deep learning (DL)

Old style multi-level NNET usually used to learn non-linear relations. Whereas DL is used to train with “stochastic gradient descent using back-propagation” as built on multi-layer “feed-forward artificial neural network.” It uses many hidden layers comprise of nodes with incorporated activation functions. DL has some advanced features, such as “L1 or L2 regularization, adaptive learning rate, momentum training, and drop out”, used to allow high predictive performance [42]. As per Li and colleagues, DL works from a neural network that offers information about other data as input and produces the outcome by using many layers [26].

DL initiates the method by using widespread hidden layers that contain nodes to produce the result, while the traditional neural network only considers a single hidden layer [55]. According to Schmidhuber, the old NNET requires more material for features to perform

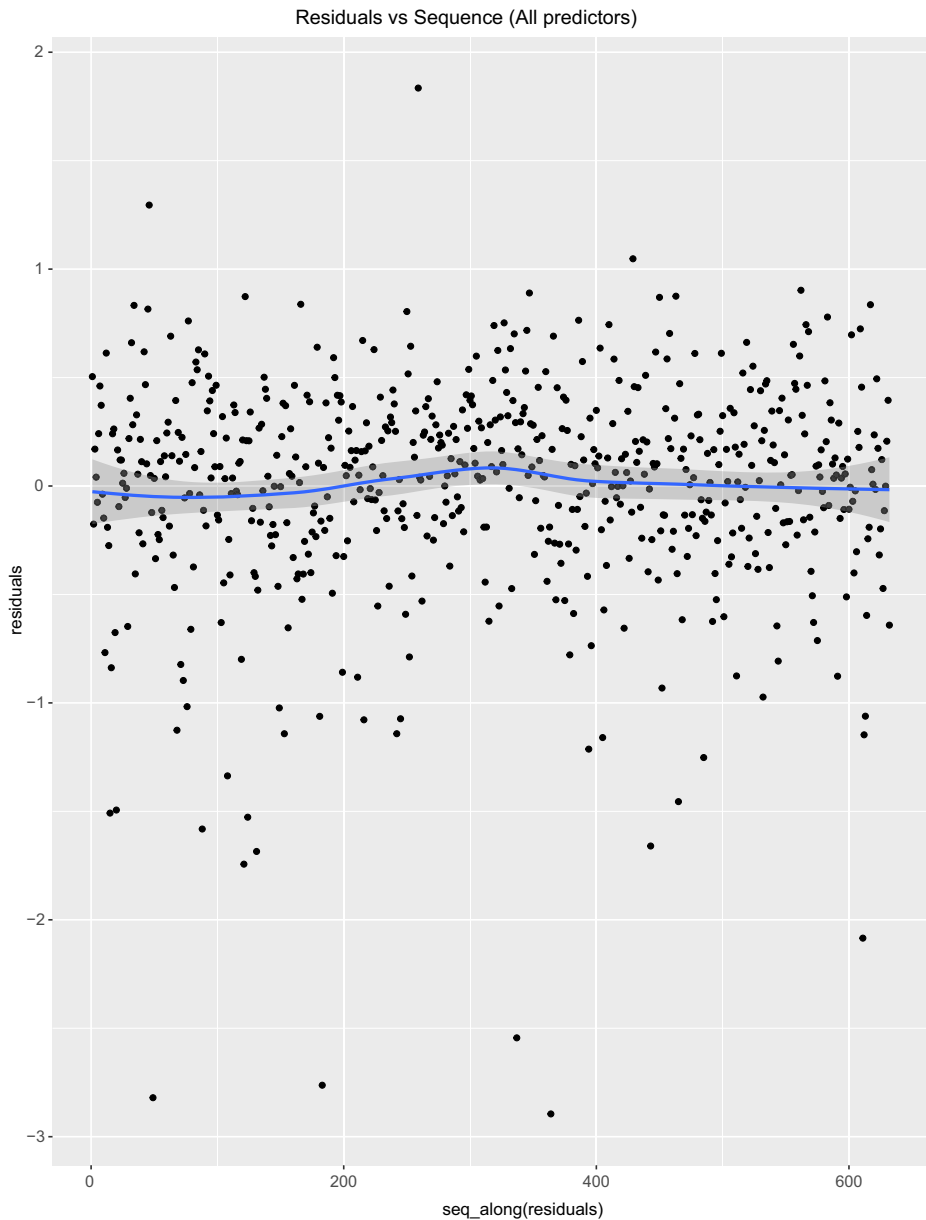


**Fig. 10** Statistical model predicted values vs. residuals

feature selection, and for domain knowledge of data. On the contrary, DL does not require any substantial facts about features [46]. Xing and Du validates that DL can automatically tune and select the model at an optimum level and also has the built-in quality to mine the features without any participation and collaboration of humans, which fabulously saves much time [55].

### 5.3 Decision tree (DT)

It is a tree-like structure, has nodes, i.e., Internal and leaf. Mostly, data whose output label is unknown, DTs are implemented to classify them, and the route from root to leaf must be trailed. It made by training data which comprise of data records, and each record formed by a



**Fig. 11** Representation of residuals vs. all predictors

set of features and output label. It covers either distinct or continuous values [21]. DT is a distinguishable and straightforward structure. Each node represents the splitting rule for a feature to classify the target value. Dataset has nominal and numerical features, DT can implement. Primarily, in DT, a response variable must be nominal for classification and numerical for regression [42].

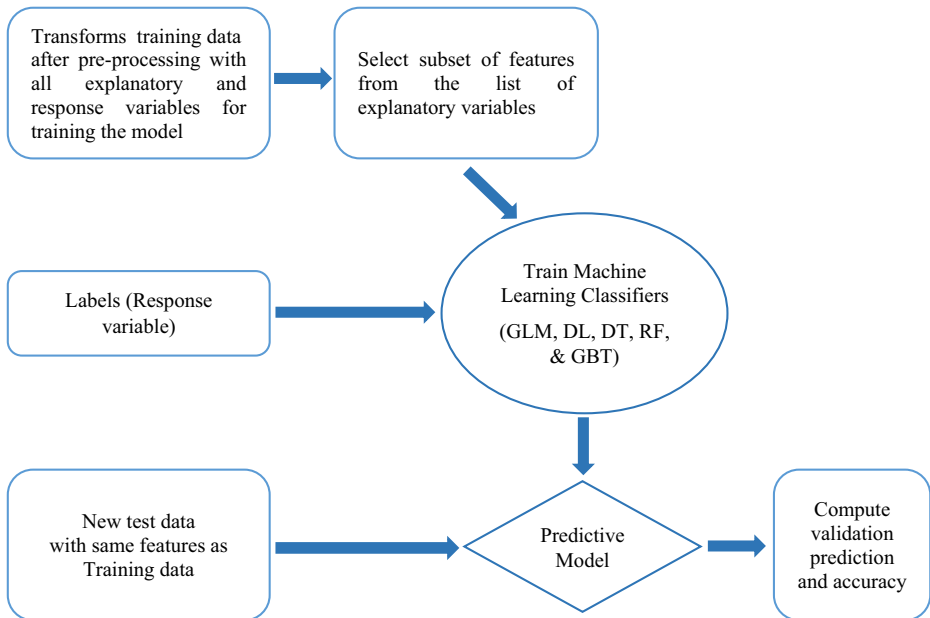


Fig. 12 Classifiers training and testing algorithm flowchart

## 5.4 Random forest (RF)

RF by Breiman links numerous tree input variables in a group. Trees could be broken down when new incidences classified, and each tree states a classification [8]. From a cumulative number of polls quantified by the group of trees, the forest then elects which label to assign to this new occurrence [1].

[31] applied this technique to predict the fortitude of students in science and engineering discipline. RF produces many random trees on different subgroups of data, and the successive model builds on the polling of these trees.

Due to this modification, it is less likely to overtraining. Minimal leaf size for the classification task is 2 and 5 for regression [42].

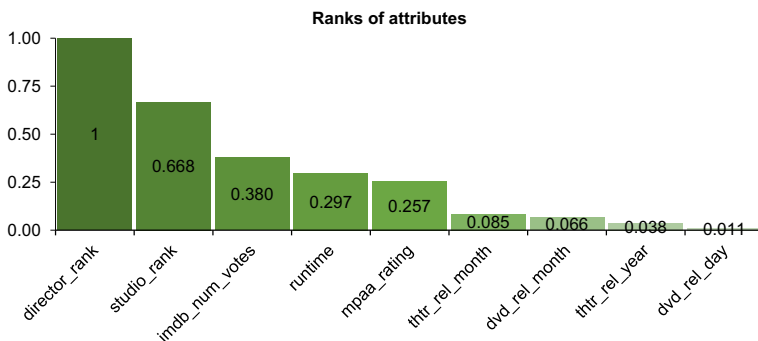


Fig. 13 Representation of ranks of the attributes



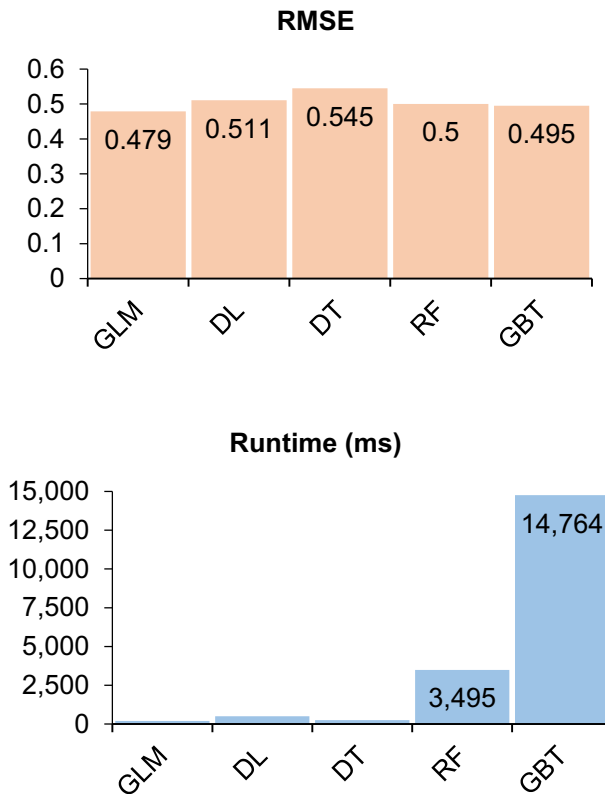


Fig. 14 Five learning models' summary of RMSE and runtime in milliseconds

RF is an algorithm that combines the arbitrarily made autonomous DTs to make predictions [8, 24]. Generally, RF presents meaningfully improved performance. Moreover, RF has an excellent capability to deal with irrelevant inputs [34].

### 5.5 Gradient boosted trees (GBT)

It has the proficiencies of parallel computing and also the active linear model solver. Due to these capabilities, it produces excellent performance and accuracy, also linked to Gradient boosted machine (GBM), another boosted algorithm. Moreover, it can form decision trees which are distinct logical models [11].

It is an ensemble of either classification or regression tree models. They are forward learning methods that attain predictive results through increasingly better estimations. By applying weak classification algorithms to gradually changing data, lots of DTs created that produce groups of weak prediction models. Though boosting trees enhances their accuracy and performance. It also reduces the speed and “human interpretability.” This process simplifies tree boosting to curtail these issues [42].

GBT executes similarly to Adaptive tree boosting (ATB), another boosting algorithm. At each iteration, it uses residuals of the last prediction function [56]. GBT uses some different measures, i.e., binomial deviance, to identify the cost of errors, and it differs from ATB [9, 18]. In the case of a

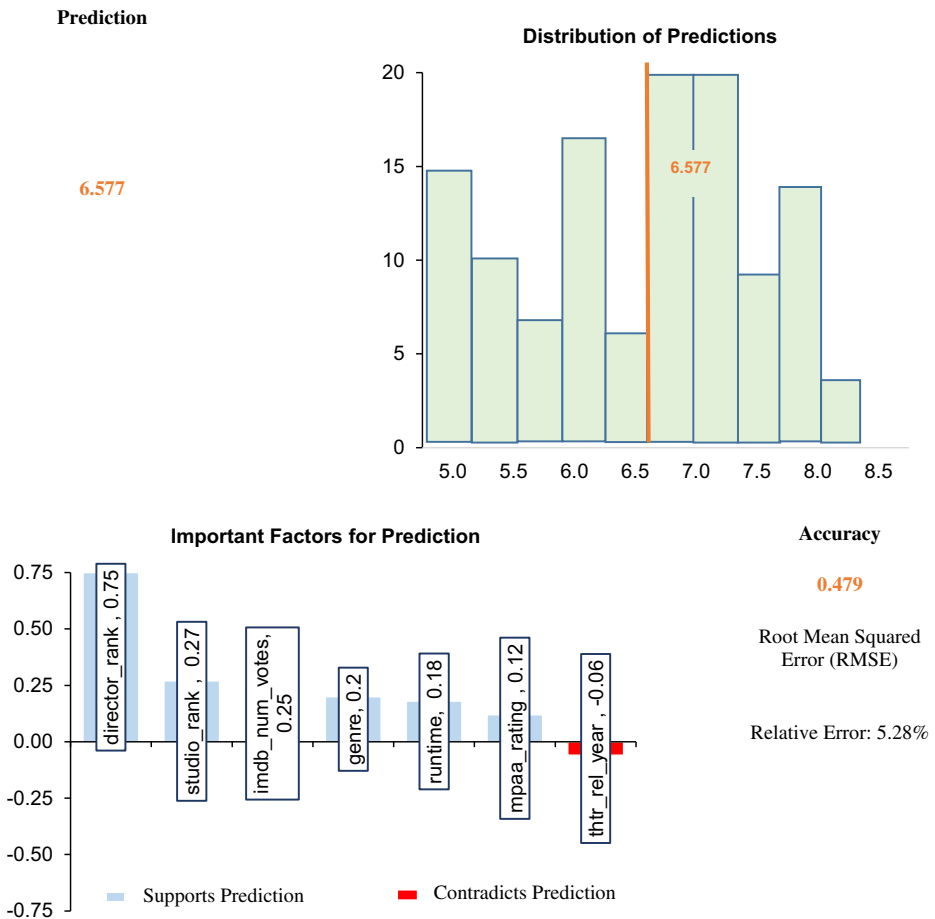


Fig. 15 High accuracy achieving GLM model

multicollinearity problem that exists amongst the features and the number of features is comparatively large to the number of data points, GBT is usually considered as robust [30, 40].

## 6 Evaluation

We have used RapidMiner 8.1 to implement the above-mentioned machine learning methods and tested them. Figure 12, reveals the flow chart specifying the movie prediction. As there are plenty of data mining/machine learning tools available and RapidMiner is one of those, and it is best suited for data mining tasks and contains a vast collection of machine learning algorithms. List of operators (such as *Blending*, *Cleansing*, *Modeling*, and *validation*) are available to perform mining of data.

This section defines the training and test data, as well as the performance measures used in experiments. The last subsection comprises results and analysis.

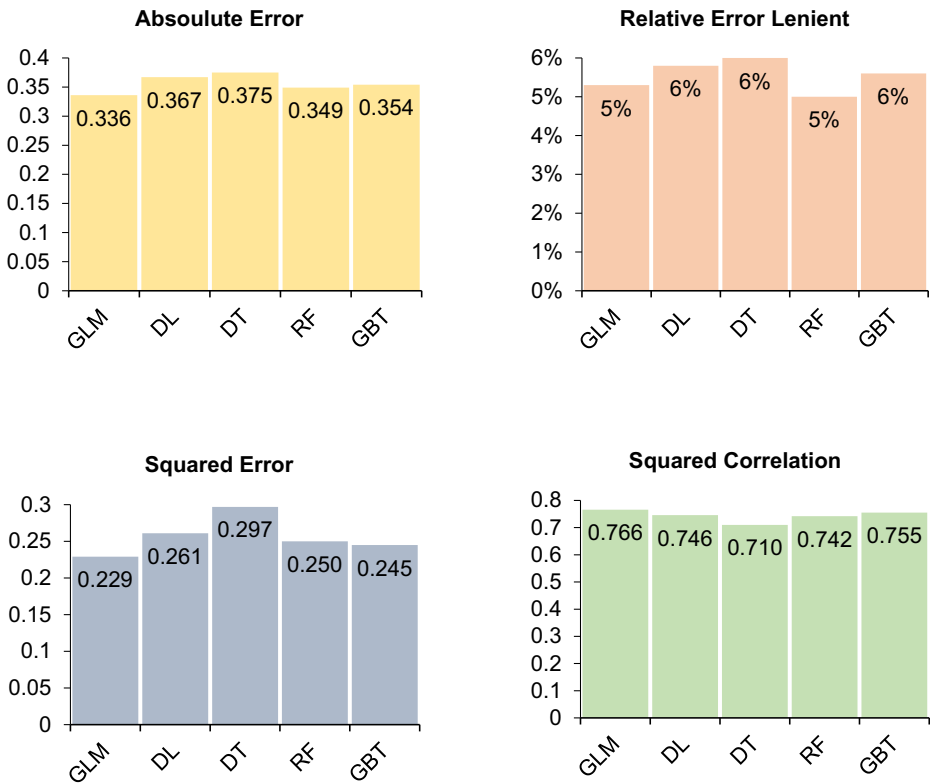


Fig. 16 Performance of five machine learning algorithms with other performance metrics

## 6.1 Training and test data

Data regarding movies and users were collected from the publicly accessible IMDb. Existing `imdb_rating` available in our data represents liking users gave in their reviews [28].

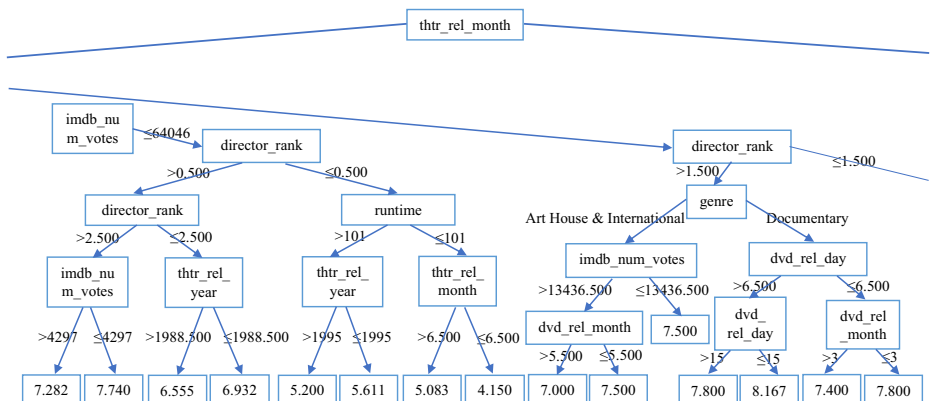


Fig. 17 Subgroups of random trees of RF

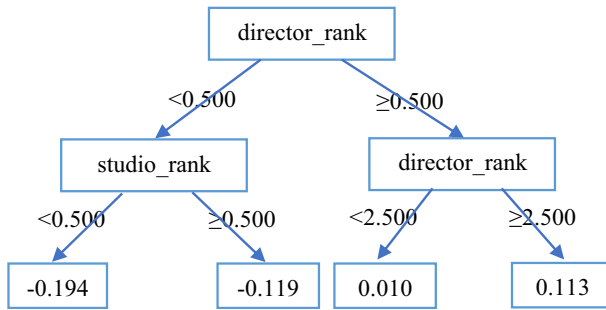


Fig. 18 Snapshot of subgroups of GBT

The training data were obtained by a repetitive random sub-sampling validation method. This technique reiterates the validation with the arbitrary partitions of training and test data. Moreover, this method resolves the k-fold cross validation issue in which, as ‘k’ grows, the size of the test data shrinks, and the performance variance of each sharp fold increases [49].

When the size of the data is small, the impact of such an issue can depreciate. Since, in this study, the size of the data set is limited, and hence, it has evidenced that repeated random sub-sampling is far better and appropriate than k-fold cross-validation [24]. So, we have split the training and test dataset into the 80:20 ratio, respectively.

## 6.2 Performance measures

In this paper, we adopted the performance metric of [28] RMSE, the most common metric used to “measure accuracy for continuous variables” (<http://yahwes.github.io/>) and also used to present the accomplishment of the numerous methods used in this study. Lower values of RMSE are better and calculated by using the equation no. 1:

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2} \quad (1)$$

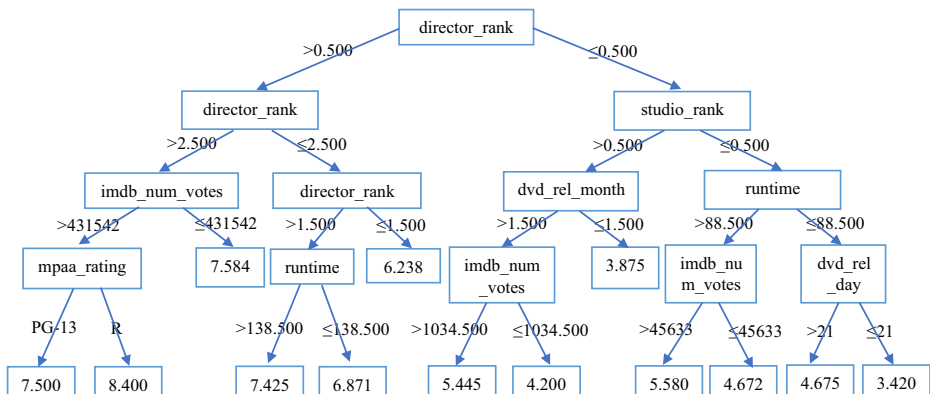


Fig. 19 DT's segregation of features

**Table 8** Important predictors for prediction

Predictors	Model				
	GLM	DL	DT	RF	GBT
director_rank	> 0.75	> 0.75	> 0.80	> 0.75	> 0.75
Studio_rank	0.25–0.30	0.25–0.30	0.0–0.3	0.22–0.24	0.7–0.9
Imdb_num_votes	0.23–0.27	0.45–0.50	–	0.23–0.27	0.48–0.52
Genre	0.18–0.22	0.10–0.15	0.5–0.10	0.12–0.15	0.9–0.12
Runtime	0.19–0.22	–	0.23–0.27	0.8–0.12	–
Mpaa_rating	0.10–0.15	0.10–0.15	0.3–0.5	0.11–0.13	0.9–0.12
Thtr_rel_year	(–0.8) – (–0.12)	(–0.1) – (–0.15)	–	–	(–0.5) – (–0.10)
Thtr_rel_month	–	–	–	0.9–0.11	–
Dvd_rel_day	–	0.10–0.15	0.5–0.10	–	0.7–0.9
Dvd_rel_month	–	–	0.3–0.4	–	–

Positive value: Supports prediction, Negative value: Contradicts prediction.

where  $n$  is the training set which contains movies related records,  $y_j$  is the real user rating for the movie in record  $j$ , and  $\hat{y}_j$  which is the predicted user rating.

We also used other performance metrics, i.e., (*Absolute error*; *Relative error*; *Squared error*; and *Squared correlation*), measured by equation nos. (2, 3, & 4).

*Absolute error (AE)* is the actual minus predicted value.

$$AE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \quad (2)$$

*Relative error (% error)* is the percentage form of AE.

$$\%error = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| * 100 \quad (3)$$

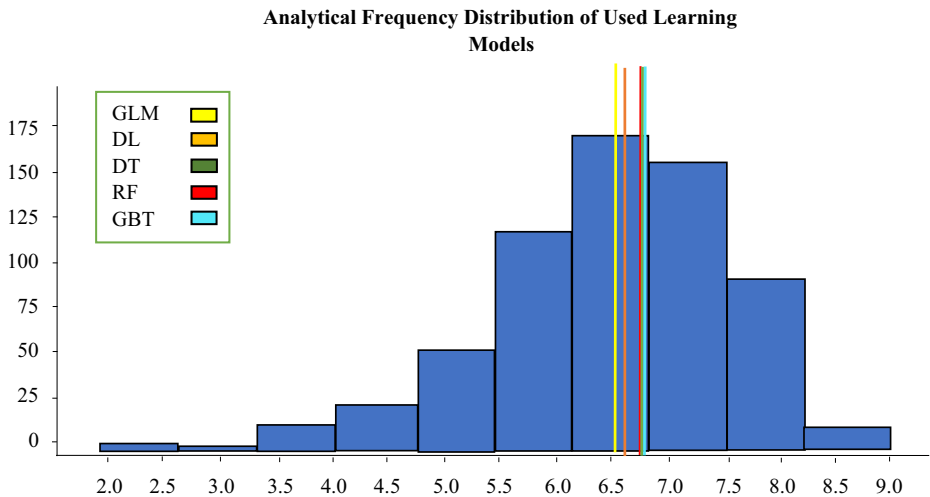
*Squared error* is the average squared difference between estimated and actual value.

$$Squared\ error = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2 \quad (4)$$

*Squared correlation ( $r^2$ )* is the square of the correlation coefficient  $r^2$ , and computed by equation no. 5. It is a useful value in linear regression and measures how close the data are to the fitted regression line. Tells us what the model explains percent of the variability in the response variable.

**Table 9** Distribution of predictions

Model	Frequency Distribution of Predictions	RMSE	Relative Error
GLM	6.577	0.479	5.28%
DL	6.693	0.511	5.77%
DT	6.871	0.545	5.97%
RF	6.868	0.500	5.49%
GBT	6.875	0.495	5.61%



**Fig. 20** Cumulative predictive frequency distribution of five machine learning methods

$$r^2 = 1 - \frac{\text{explained variability}}{\text{total variability}} \quad r^2 = 1 - \frac{\text{Residual sum of squares}}{\text{total sum of squares}} \quad r^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (5)$$

### 6.3 Results and analysis

We have implemented *five learning algorithms* and the results for each of the applied methods shown concerning their runtime in Fig. 14. Figure 13 is the weights (ranks) of the attributes which show the universal significance of each attribute for the value of the target attribute, independent of the modeling algorithm.

The RMSE has shown for every method. The picture shows the model built on machine learning techniques and methods. It also depicts the performance achieved by each regression classifier and reveals the accuracy performance by repetitive random sampling validation technique, in which it randomly reproduces division of training and test data — additionally, the results for each of the implemented methods shown in Table 7.

**Table 7** Result of machine learning methods concerning performance measures

Methods	Runtime	Measures				
		RMSE	AE	RE	SE	SC
GLM	58 ms	0.479	0.336	5.3%	0.229	0.766
DL	469 ms	0.511	0.367	5.8%	0.261	0.746
DT	152 ms	0.545	0.375	6.0%	0.297	0.710
RF	3 s	0.500	0.349	5.5%	0.250	0.742
GBT	15 s	0.495	0.354	5.6%	0.245	0.755

ms: Milliseconds, RMSE: Root mean squared error, AE: Absolute error, RE: Relative error, SE: Squared error, SC: Squared correlation

The above results demonstrate the ratio of time we can predict the cases suitably. We achieved maximum accuracy with *GLM*: 0.479, *RF*: 0.50, and *GBT*: 0.495, respectively, and lower values of RMSE are always better. The other classifiers also attained good results, i.e., *DL*: 0.511, and *DT*: 0.545.

The *GLM* model is the one with the highest accuracy among the candidate models and the best model in this study due to the lowest RMSE value as it works on the MLE principle. This algorithm fits generalized linear models to the data by maximizing the log-likelihood. The elastic net penalty can be used for parameter regularization. The model fitting computation is parallel, extremely fast, and scales extremely well for models with a limited number of predictors with non-zero coefficients. It performs parallel computation with predictors. It was trained using the regularization and split the data into an 80:20 ratio with shuffled sampling because it builds random subsets of the training set in the performance parameter, and selected the example weights. This parameter allows example weights to be used for statistical performance calculations if possible. This parameter does not affect if no attribute has a weight role. Figure 15, illustrates the summary of the learning model.

### 6.3.1 Interpretation

The prediction of the model is 6.577. Essential factors for prediction show, which predictor occurrence is of utmost significance for prediction, and in this case, the most prominent support is coming from *director\_rank*. The RMSE of all predictions done by this model is 0.479, and the relative error is about 5.28%. Also, Table 8 reveals the list of predictors according to the importance of the prediction. Figure 16, displays the other performance metrics, which we discussed in subsection 6.2. It shows the performance of five machine learning algorithms about an *Absolute error*, *Relative error*, *Squared error*, and *Squared correlation*.

It has been noted that by comparing the other performance metrics, GLM, RF, and GBT are still considered high achieving accuracy models, and GLM maintained high accuracy performance, attained above 76% squared correlation.

Furthermore, RF and GBT are an ensemble of arbitrarily made autonomous DTs and an ensemble of classification or regression tree models. Figs. 17 and 18, exhibits the snapshot of random trees on different subgroups of data. Figure 19, demonstrates the DT structure, which is distinct and straightforward. Finally, Fig. 20, validates the overall prediction distribution of five machine learning models used in this study, and exhibits all the learning models are nearly normally distributed shown in Table 9.

*Random forest (RF)*: It works on the bagging techniques as it is the combination of trees which randomly selects predictors at each possible split. It creates the bootstrapped dataset that is the same size as the original dataset and selects the random samples from the original dataset and then creates DTs using the bootstrapped dataset but only use a random subset of (variables) features (or column) at each step. As in this study, we are dealing with regression, so, in the end, the leaf nodes showing the known prediction values, i.e., *imdb\_rating*.

*Gradient boosted tree (GBT)*: It also works on the ensembles of DTs and typically works on the boosting algorithm, which converts the weak prediction into the strong predictions. Moreover, it is different from the RF, as RF uses DTs, whereas GBT uses regression trees for prediction, as our predicted outcome is real no., i.e., *imdb\_rating*.



*Decision tree (DT)*: It selects all the features or columns from the entire dataset and then picks up one feature as a root. However, the question is how to pick the first attribute at the root? The answer is it selects according to the values given for that attribute or feature and compares & counts, which has the higher votes.

## 7 Conclusion and future works

This study demonstrates to predict the popularity of a movie. We have implemented a machine learning approach along with the statistical modeling for our investigation. Machine learning has plenty of robust algorithms for classification and regression. The primary objective of this research is to improve and compare the previous researches. As a result, after performing the regression, our model has predicted the popularity of the movie with the accuracy performance in terms of squared correlation (SC) is GLM: 76.6%, DL: 74.6%, DT: 71.0%, RF: 74.2%, and GBT: 75.5%, respectively. The features that contributed the most significant support are from *director\_rank*, *studio\_rank*, *genre*, *runtime*, *mpaa\_rating*, *imdb\_num\_votes*, and *dvd\_rel\_month*. Moreover, the essential support is coming from *director\_rank*, which is considered to be an important factor/predictor for prediction shown in Table 8 above, and it also confirms that the director feature is the most significant attribute for the popularity of the movie, and must be taken into consideration.

Furthermore, it is hard to perform data mining on IMDb due to lots of attributes relating to a movie in a variable scope. Our study has many moral implications, both statistically and practically. To our knowledge, our research, amongst the previous studies, is one of the few studies that have focused on the feature aspect. We have chosen features based on statistical techniques and criteria. Most of the forecasted studies using machine learning techniques emphasize on the augmentation of the predictive power, means they only focus on the building of better performing model irrespective of the model's features taking into consideration for a better outcome. It raises a question on the black-box nature of the machine learning techniques. However, by identifying what features to include based on statistical theories, we can defend such negative reviews and criticism.

The predictive model presented here may be used to predict *imdb\_rating* for a movie. It should be noted that the model based on a tiny sample, and some studios and directors were not sufficiently represented in the data set, which may decrease the usefulness of the model for these particular types of movies. Another shortcoming is the limited number of variables that we were able to retain in our final model. A more extensive training set with additional features are the key aspects and may improve the overall performance of the model.

We foresee our future research on the popularity of the movie in three main directions. First, we would like to experiment with a few approaches that are adequate optimization parameters, and criterion can be considered to improve and increase the accuracy of our model. Second, though, machine learning methods implemented in this study are entirely appropriate and comprehensive. However, still, many techniques can be explored and applied to solve the prediction problem in the movie domain. Third, other features could be incorporated to construct a more accurate model. We suppose that these recommendations could improve the prediction accuracy of movie popularity.

**Acknowledgments** The effort of this paper supported by “NATIONAL NATURAL SCIENCE FOUNDATION OF CHINA, grant number 91630206, and 61572434”, and “THE NATIONAL KEY R&D PROGRAM OF CHINA, grant number 2017YFB0701501”.

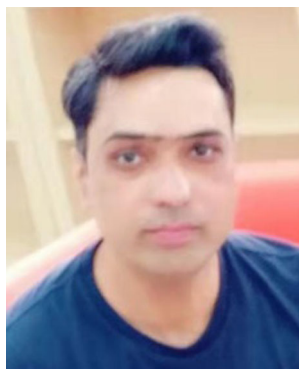
## References

1. Aguiar E, Chawla NV, Brockman J et al (2014) Engagement vs performance. *Proceedings fourth Int Conf learn anal Knowl - LAK '14* 103–112. <https://doi.org/10.1145/2567574.2567583>
2. Asad KI, Ahmed T, Saiedur Rahman M (2012) Movie popularity classification based on inherent movie attributes using C4.5, PART and correlation coefficient. *2012 Int Conf informatics. Electron Vision, ICIEV 2012*:747–752. <https://doi.org/10.1109/ICIEV.2012.6317401>
3. Asur S, Huberman BA (2010) Predicting the future with social media. *Web Intell Intell agent Technol (WI-IAT)*, 2010 IEEE/WIC/ACM Int Conf on, Vol 1 IEEE 1:492–499
4. Asur S, Huberman BA (2010) Predicting the future with social media. *Proc - 2010 IEEE/WIC/ACM Int Conf web Intell WI 2010* 1:492–499. <https://doi.org/10.1109/WI-IAT.2010.63>
5. Babu SP (2014) Predicting movie success based on IMDB data. *Int J Data Min Tech Appl Integr Intell Res* 03:365–368
6. Basuroy S, Chatterjee S, Ravid SA (2003) How critical are critical reviews? The box office effects of film critics, star power, and budgets. *J Mark* 67:103–117. <https://doi.org/10.1509/jmkg.67.4.103.18692>
7. Billsus D, Pazzani MJ (1998) Learning collaborative information filters. *Proc Fifteenth Int Conf Mach Learn* 54:48
8. Breiman L (2001) Random forests. *Mach Learn* 45:5–32
9. Chambers M, Dinsmore TW (2015) *Advanced Analytics Methodologies: Driving Business Value with Analytics*:324
10. Cizmeci B, Oguducu SG (2018) Predicting IMDb ratings of pre-release movies with factorization machines using social media. *UBMK 2018 - 3rd Int Conf Comput Sci Eng* 173–178. <https://doi.org/10.1109/ubmk.2018.8566661>
11. Cobos R, Wilde A, Zaluska E (2017) Predicting attrition from massive open online courses in FutureLearn and edX. *CEUR Workshop Proc* 1967:74–93
12. De Vany A, Walls WD (1999) Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office? De Vany, Arthur, and W. David Walls. "Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office?" *Journal of Cultural Economics*, vol. 23, n. J Cult Econ 23:285–318. <https://doi.org/10.1023/A:1007608125988>
13. Du J, Xu H, Huang X (2014) Box office prediction based on microblog. *Expert Syst Appl* 41:1680–1689. <https://doi.org/10.1016/j.eswa.2013.08.065>
14. Elberse A (2008) The power of stars: do star actors drive the success of movies? *J Mark* 71:102–120. <https://doi.org/10.1509/jmkg.71.4.102>
15. Eliashberg J, Hui SK, Zhang ZJ (2007) From story line to box office: a new approach for green-lighting movie scripts. *Manag Sci* 53:881–893. <https://doi.org/10.1287/mnsc.1060.0668>
16. Gallagher J (2008) Netflix case study: David becomes goliath. *Gall com*:1–16
17. Han J, Kamber M (2004) *Data mining concepts and techniques*. Morgan Kaufman Publ
18. Hastie T, Tibshirani R, Friedman J (2017) *The elements of statistical learning; data mining, inference, and prediction*. Second Ed 757. <https://doi.org/10.1007/b94608>
19. Im D, Nguyen MT (2011) Predicting box-office success of movies in the U . S . Market. *Cs* 1–5
20. Ishikawa M, Geczy P, Izumi N, et al (2007) Information diffusion approach to cold-start problem. *Proc - 2007 IEEE/WIC/ACM Int Conf web Intell Intell agent Technol - work WI-IAT work 2007* 129–132. <https://doi.org/10.1109/WIIATW.2007.4427556>
21. Kabra RR, Bichkar RS (2011) Performance prediction of engineering students using decision trees. *Int J Comput Appl* 36:8–12
22. Kim Y, Kang M, Jeong SR (2018) Text mining and sentiment analysis for predicting box office success. *KSII Trans Internet Inf Syst* 12:4090–4102. <https://doi.org/10.3837/tiis.2018.08.030>
23. Latif MH, Afzal H (2016) Prediction of movies popularity using machine learning techniques. *IJCSNS Int J Comput Sci Netw Secur* 16:127–131
24. Lee K, Park J, Kim I, 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>
25. Lew MS, Sebe N, Djeraba C, Jain R (2006) Content-based multimedia information retrieval: state of the art and challenges. *ACM Trans Multimed Comput Commun Appl* 2:1–19. <https://doi.org/10.1145/1126004.1126005>
26. Li W, Gao M, Li H, et al (2016) Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning. *Proc Int Jt Conf Neural Networks 2016-Octob*:3130–3137. doi: <https://doi.org/10.1109/IJCNN.2016.7727598>
27. Litman BR (1983) Predicting success of theatrical movies: an empirical study. *J Pop Cult* 16:159–175. [https://doi.org/10.1111/j.0022-3840.1983.1604\\_159.x](https://doi.org/10.1111/j.0022-3840.1983.1604_159.x)
28. Marovic M, Mihokovic M, Miksa M, et al (2011) Automatic movie ratings prediction using machine learning. *2011 Proc 34th Int Conv MIPRO* 1640–1645

29. Masih S, 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
30. Mayr A, Binder H, Gefeller O, Schmid M (2014) The Evolution of Boosting Algorithms From Machine Learning to Statistical Modelling \*. 1–32
31. Mendez G, Buskirk T, Lohr S, Haag S (2008) Factors associated with persistence in science and engineering majors: an exploratory study using classification trees and random forests. *J Eng Educ* 97:57
32. Mestyán M, Yasseri T, Kertész J (2013) Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data PLoS One:8. <https://doi.org/10.1371/journal.pone.0071226>
33. Mishne G, Glance N (2005) Predicting movie sales from blogger sentiment. AAAI Spring Symp Comput Approaches to Anal Weblogs:155–158. <https://doi.org/10.1016/j.cger.2010.02.002>
34. Montillo AA (2009) Statistical foundations of data analysis. Springer, New York
35. Nelson RA, Glotfelty R (2012) Movie stars and box office revenues: an empirical analysis. *J Cult Econ* 36: 141–166. <https://doi.org/10.1007/s10824-012-9159-5>
36. Ng VKY, Cribbie RA (2018) The gamma generalized linear model, log transformation, and the robust Yuen-Welch test for analyzing group means with skewed and heteroscedastic data. *Commun Stat Simul Comput* 0918:1–18. <https://doi.org/10.1080/03610918.2018.1440301>
37. Oghina A, Breuss M, Tsagkias M, De Rijke M (2012) Predicting IMDB movie ratings using social media. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 7224 LNCS: 503–507. doi: [https://doi.org/10.1007/978-3-642-28997-2\\_51](https://doi.org/10.1007/978-3-642-28997-2_51)
38. Popescul A, Pennock DM, Lawrence S (2001) Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. *Proc Seventeenth Conf Uncertain Artif Intell* 2001:437–444
39. Prag J, Casavant J (1994) An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. *J Cult Econ* 18:217–235. <https://doi.org/10.1007/BF01080227>
40. Prettenhofer P (2014) Louppe G (2014) gradient boosted regression trees in Scikit-learn. In *PyData, London*
41. Quader N, Gani MO, Di C (2018) Performance evaluation of seven machine learning classification techniques for movie box office success prediction. In: *3rd Int Conf Electr Inf Commun Technol EICT 2017* 2018-January, pp 1–6. <https://doi.org/10.1109/EICT.2017.8275242>
42. RapidMiner (2016) RapidMiner Documentation. <https://docs.rapidminer.com/latest/studio/operators/>.
43. Rundel MC (2018) Linear Regression and Modeling. <https://www.coursera.org/learn/linear-regression-model>.
44. Sarwar B, Karypis G, Konstan J, Riedl J (2000) Analysis of recommendation algorithms for e-commerce. *Proc 2nd ACM Conf Electron Commer - EC '00* 158–167. <https://doi.org/10.1145/352871.352887>
45. Schein AI, Popescul A, Ungar LH, Pennock DM (2002) Methods and metrics for cold-start recommendations. *Proc 25th Annu Int ACM SIGIR Conf res Dev Inf Retr - SIGIR '02* 253. <https://doi.org/10.1145/564376.564421>
46. Schmidhuber J (2015) Deep learning in neural networks. *Neural Networks* 61:85–117. doi: <https://doi.org/10.1016/j.neunet.2014.09.003>
47. Sharda R, Delen D (2006) Predicting box-office success of motion pictures with neural networks. *Expert Syst Appl* 30:243–254. <https://doi.org/10.1016/j.eswa.2005.07.018>
48. Simonoff JS, Sparrow IR (2015) Predicting movie grosses: winners and losers, blockbusters and sleepers. *Chance* 13:15–24. <https://doi.org/10.1080/09332480.2000.10542216>
49. Smith MR, Mitchell L, Giraud-Carrier C, Martinez T (2014) Recommending learning algorithms and their associated hyperparameters. *CEUR Workshop Proc* 1201:39–40. <https://doi.org/10.1145/2487575.2487629>
50. Son J, Kim SB (2017) Content-based filtering for recommendation systems using multiattribute networks. *Expert Syst Appl* 89:404–412. <https://doi.org/10.1016/j.eswa.2017.08.008>
51. Tang TY, Winoto P, Guan A, Chen G (2018) “The foreign language effect” and movie recommendation: a comparative study of sentiment analysis of movie reviews in Chinese and English. *ACM Int Conf proceeding Ser* 79–84. <https://doi.org/10.1145/3195106.3195130>
52. Vu DH, Muttatqi KM, Agalgaonkar AP (2015) A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables. *Appl Energy* 140:385–394. <https://doi.org/10.1016/j.apenergy.2014.12.011>
53. Wang H, Zhang H (2018) Movie genre preference prediction using machine learning for customer base information. 110–116. <https://doi.org/10.1109/CCWC.2018.8301647>
54. Wilson DC, Smyth B, Sullivan DO (2003) Sparsity reduction in collaborative recommendation: a case-based approach. *Int J Pattern Recognit Artif Intell* 17:863–884. <https://doi.org/10.1142/s0218001403002678>
55. Xing W, Du D (2018) Dropout prediction in MOOCs: using deep learning for personalized intervention. *J Educ Comput Res*. <https://doi.org/10.1177/0735633118757015>
56. Yamagishi J, Kawai H, Kobayashi T (2008) Phone duration modeling using gradient tree boosting. *Speech Commun* 50:405–415. <https://doi.org/10.1016/j.specom.2007.12.003>
57. Yu L, Liu L, Li X (2005) A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-commerce. *Expert Syst Appl* 28:67–77. <https://doi.org/10.1016/j.eswa.2004.08.013>

58. Zhang W, Skiena S (2009) Improving movie gross prediction through news analysis. Proc - 2009 IEEE/WIC/ACM Int Conf web Intell WI 2009 1:301–304. <https://doi.org/10.1109/WI-IAT.2009.53>
59. Zhang L, Luo J, Yang S (2009) Forecasting box office revenue of movies with BP neural network. Expert Syst Appl 36:6580–6587. <https://doi.org/10.1016/j.eswa.2008.07.064>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Syed Muhammad Raza Abidi** is a doctoral candidate in the School of Computer Engineering and Science, Shanghai University, Shanghai, China. He has received his MIT (Master of Information Technology) degree from Virtual University of Pakistan in 2014. His research has been published in *Sustainability (MDPI Journal)*, *Artificial Intelligence Review (Springer Journal)*, *Computational Intelligence and Neuroscience (Hindawi Journal)*, and accepted in various conferences. His research interest covers data analytics, big data, and educational data mining special focus on intelligent tutoring systems, Massive Open Online Courses (MOOCs), and movie industry based on machine learning approach. E-mail: [razaabdi@live.com](mailto:razaabdi@live.com), [abidi@i.shu.edu.cn](mailto:abidi@i.shu.edu.cn)



**Yonglin Xu** is a master student in the School of Computer Engineering and Science, Shanghai University, Shanghai, China. He has received his BSCS (Bachelor of Science in Computer Science) degree from Beijing Union University, Beijing, China. His current research is focused on application of data mining/machine learning in materials science, and educational data mining. E-mail: [wffzxy1@gmail.com](mailto:wffzxy1@gmail.com)



**Jianyue Ni** is a PhD scholar in the School of Computer Engineering and Science, Shanghai University, Shanghai, China. He earned his MSc. (Master of Science in Advanced Software Engineering) degree from University of Sheffield, England. His research interests are in the areas of Machine learning, Materials design, and High-performance computing. E-mail: njy0612@163.com



**Xiangmeng Wang** is a master student in the School of Computer Engineering and Science, Shanghai University, Shanghai, China. She has received BS (Bachelor of Science in Educational Technology) degree from China West Normal University, Nanchong, Sichuan, China. Her current research is focused on application of data mining/machine learning in materials science, and educational data mining. E-mail: xiangm\_wang@163.com



**Dr. Wu Zhang** is a Professor and department head in the School of Computer Engineering and Science, Shanghai University, Shanghai, China. He earned his PhD (Doctor of Philosophy) degree from Northwestern Polytechnic University, California, USA. His main research fields include: Application of machine learning on materials science, high performance computing and applications, data driven decision modelling based on machine learning approach, and self-adaption and parallel algorithm, scientific and engineering computing and its application software, bioinformatics, and fluid dynamics. His refereed articles have appeared in various academic journals and conference proceedings. E-mail: wzhang@shu.edu.cn

## Affiliations

**Syed Muhammad Raza Abidi<sup>1</sup> • Yonglin Xu<sup>1</sup> • Jianyue Ni<sup>1</sup> • Xiangmeng Wang<sup>1</sup> • Wu Zhang<sup>1,2</sup>**

Yonglin Xu  
wffzxy1@gmail.com

Jianyue Ni  
njy0612@163.com

Xiangmeng Wang  
xiangm\_wang@163.com

<sup>1</sup> School of Computer Engineering and Science, Shanghai University, No. 99, Shangda Road, Baoshan Campus, Baoshan District, Shanghai 200444, China

<sup>2</sup> Shanghai Institute of Applied Mathematics and Mechanics, Shanghai University, No. 99, Shangda Road, Baoshan Campus, Baoshan District, Shanghai 200444, China