J. Management and Humanity Research
Vol. 8, 2022, 57-75
ISSN: 2582-7766 *(online)*
Published on 31 December 2022
www.researchmathsci.org
DOI:http://dx.doi.org/10.22457/jmhr.v08a062254

Journal of
Management and
Humanity Research

A Machine Learning-Driven Movie Performance Prediction System to Improve Decision-Making Capability of Movie Investors

Chiranjib Paul[1*] **and *P. K. Das***[2]

[1,2]Indian Institute of Foreign Trade, Kolkata, West Bengal – 700100, India.
[2]E-mail: pkdas@iift.edu
[*]Corresponding author. [1]E-mail: chiranjib_2002@yahoo.com,

Received 2 December 2022; Accepted 30 December 2022

Abstract. Moviegoers refer to online audience movie ratings before deciding to watch a movie. They are more inclined to watch a movie with a high average rating. We develop a system to predict average audience movie ratings based on the lead cast and crew at an early stage of movie production. After valuing multiple scenarios, investors can use our study to select the lead cast and crew objectively. Judicious selection of the key cast and crew is extremely important as investors commit to large sums of money as professional fees while signing contracts with them. Our study uses a relatively large sample of 1687 Indian movies spread across 10+ languages released in India between 2010 and 2019 to identify the important predictors influencing average audience movie rating. Identification of important predictors improves the explainability of the prediction model, which increases the investors' trust in the predicted values. The best model, random forest, reduces the baseline prediction error of the average rating by 10.21%.

Keywords: Movie audience rating prediction; Support Vector Regression; Artificial Neural Network; Random Forest; Decision support system; Indian movies

1. Introduction

Movie investment is a risky business [1,2]. Investors make a large financial commitment in the initial phase during the selection of the lead actors, actresses, and important crew members like movie directors, music composers, producers, and writers, whereas the revenue starts pouring in after the release of the movie. There have been numerous examples when the lead cast and crew fail to deliver desired movie success [3,4], even though these lead stars and crew members do charge a hefty fee [5]. Therefore, movie investors prefer to assess the potential performance of the movie depending on the selection of the lead cast and crew. The objective of our study is to develop a decision support system for movie investors to predict movie performance for multiple scenarios with different combinations of the lead cast and crew and movie genre. Another important aspect of our study is to predict movie performance at the early stage of the movie production rather than at the end stage, either just before or after the release of the movie. Movie performance prediction after the release of the movie is more accurate but is not

useful for investors as very limited options are available to the producer after release. Therefore, investors prefer a decision support system that can predict movie performance at the early stage, so that they have sufficient time to take corrective actions.

The study by Dastidar and Elliott [6] and Elliott et al. [7] found that the average online viewer rating (AOVR) is a significant predictor of movie revenue. Hean et al. [8] established that movie audiences are more inclined to watch a movie, that receives higher AOVR. Given the strong positive association between movie revenue and the AOVR, investors like to accurately predict AOVR at the early stage of movie production. In contrast to using movie revenue as a popular representation of movie performance [6,9], our study uses AOVR to represent movie performance. Higher AOVR represents better movie performance.

The past studies evaluated the quality of the lead cast and crew based on 1) the star power index published by film magazines [6,10], 2) the inclusion of stars in a master list [11], 3) winners of awards [5,12], 4) professional critic rating [13]. Our study, in contrast, explores the quality of the lead cast and crew of the current movie based on the performance of their past movies over the short-term, medium-term, and long-term.

Our study is conducted on Indian movies. The primary reason for the selection of Indian movies is due to limited academic research on Indian movies. This is surprising given the fact that India produces the maximum number of movies in the world [14,15,16]. The replication of studies done earlier primarily in the Hollywood context is not suggested in the Indian context due to the socio-cultural aspects of Indian moviegoers, which is accentuated by the term “Indian Touch” [17]. The viewers' interaction with the film on-screen is significantly different from the viewers of Hollywood movies in the US. Respectful silence is not at all integral to the way Indians express their appreciation of cinema [17]. Baz Luhrmann, director of Moulin Rouge said that watching a Bollywood film in Rajasthan represented a seminal moment in his understanding of cinema [17]. Srinivas [18] also highlighted the overtly participatory and interactive style of Indian audiences, a phenomenon that has eluded study in Western societies. Indian movies are structurally different from Hollywood movies [6,17,19] because 1) they contain multiple songs and dances 2) they have longer screening time, and 3) they are melodramatic. Our study considers the music composer as one of the key crew members due to the importance of music and dance in Indian movies.

A study of multiple reports in leading Indian newspapers [20,21,22], reveals that the largest contributor to the total cost of moviemaking in India is the fees charged by leading actors and actresses which on average contribute to 40-50% of the total movie budget. Hence, Indian movie investors need to predict movie performance based on the selection of key cast and crew much earlier in the movie production.

The machine learning (ML) based prediction models developed in our study assists the investors to predict AOVR more accurately depending on the selection of key cast and crew and the movie genres. This study helps the investors to finalize the lead cast and crew objectively after evaluating multiple alternatives at an early stage. The transparent reasoning (comparison of AOVR between multiple scenarios) of selecting one scenario over multiple other scenarios expedites the adoption of our proposed solution [23]. Our solution also provides explainability to movie investors by highlighting the relative importance of cast and crew for higher AOVR. Providing explainability enhances investors' trust [24] in the predicted AOVR value.

The following section (Related Studies) depicts the previous academic work in the field of forecasting movie performance (both revenue and rating), the impact of key cast and crew on movie performance, Influence of social media on the movie. The subsequent sections delineate the data collection process and the analysis methodology. The findings follow the methodology section. We conclude our study with management discussions and research limitations.

2. Related studies

2.1. Forecasting algorithms

There are three major categories of forecasting algorithms used in movie revenue prediction. These are 1) statistical learning-based models 2) diffusion-based models, and 3) ML models. The most used statistical learning-based model is Multiple Linear regression [13,25]. The advantage of the approach is its simplicity and the ability to quantify the impact of each independent variable on the target variable. The accuracy of this model has been lower due to linearity constraints. Diffusion-based models have received prominence within time-series-based forecasting models. The objective of the diffusion approach is to analyze the acceptance of a new product or/and service by the customers [26]. In a study, Jedidi et. al. [27] distributed 102 movies into four different clusters using the exponential decay model. The clusters were formed using variables like star power, MPAA rating, genre, competition, awards, seasonality, etc. Dellarocas et al. [28] introduced Bass Diffusion Model to forecast box office revenue. Lee et al. [29] used a generalized Bass model using multiple seasonal factors and herding effects, along with internal and external influencers to forecast box office revenue. In recent years, academic studies have been using ML algorithms to forecast box office performance. Most of the studies have designed it as a classification problem where forecasting is done to determine whether a movie is likely to earn higher or lower than a certain revenue value instead of designing it as a regression model to predict earnings. Delen et al. [30] developed classification-based forecasting models using discriminant analysis, decision trees, and ANN. Zhang et al. [31] also developed a classification (six predefined categories) based prediction model using ANN as a base model. Lee and Chang [32] developed a classification-based (three categories) forecasting model using a Bayesian belief network. In contrast, Abel et al. [33] built a regression-based forecasting model by applying 8 ML algorithms.

2.2. Relevance of social media and online movie ratings

In the last decade or more, another new trend has emerged in the field of marketing, which in turn affects the movie market also. Marketers are increasingly using social media platforms to connect and engage consumers. These platforms have a direct influence on movie performance through electronic word of mouth (eWoM) generated through online reviews, blogs, micro-blogging sites, and online communities. Analysis of the previous research articles [2,25,34, 35] highlighted the use of WoM sourced from the social media websites like Twitter, and Facebook, and review comments from movie databases like IMDB, Rotten Tomatoes for movie revenue prediction. As a continuation of the trend of sharing comments and opinions about movies on social media, moviegoers also started providing movie ratings on popular review-aggregation websites like IMDB, Rotten

Tomatoes, and other similar movie websites across the world. The proliferation of online review aggregation platforms effectively connects potential movie audiences and has a great effect on the dynamics of the WoM [36,37]. Researchers conducted studies to assess the impact of online ratings on movie revenue [6,13].

2.3. Movie rating prediction

It is surprising to observe that, movie rating is used to predict movie revenue, but there are very few studies, which focus on predicting movie rating. The study by Moon et al. [38] analyzed the dynamic effect between movie revenue and movie rating and found that movies with higher ratings increased revenue, and vice versa. The study by Liu et al. [39] built a system to predict movie ratings based on the sentiment of review- summarization. The source of the data was the movie reviews from Internet blogs without any numerical rating information. Ratings of movies were done based on sentiment scores determined by semantic orientation. Schmit & Wubben [40] predicted the rating of a movie from tweeter content. They used different combinations of n-gram techniques and created features using TF-IDF vectors. Multiple ML algorithms were used on the features to predict movie ratings. The study by Oghina et al. [41] predicted movie ratings using social media data. The study extracted two different sets of features, the combination of surface feature (fraction of likes per dislikes sourced from Youtube) and textual feature (text content) sourced from Twitter, generated the best rating prediction model.

The limitation with rating prediction using social media comments lies in the short-term nature of prediction. Movie investors have very limited wherewithal to change the outcome of the movie, when the rating prediction is known, which is just before or after the release of the movie. Our objective is to predict movie ratings during the early stage of movie production. The only study [42] which matches our objective developed a rating prediction model using data available during the pre-production stage. The study did not use any social media comments, instead used 1) budget 2) genre, 3) MPAA rating, 4) movie duration, 5) release date, 6) Facebook likes of the director, the top three leading actors, and cast, 7) box office performance and rating of the previous movie of the leading actors and director participated in the current movie. The study used the movie budget as one of the predictors. Sourcing reliable budget information across many movies is extremely challenging [2]. Moreover, information related to the movie production budget of Indian movies is unreliable [43]. Therefore, our study excludes budget as a predictor. The same study calculated the historical power of the lead cast and crew based on box office performance and rating of only the previous movie in which the lead cast and crew of the current movie participated. Our study has evaluated the power of the lead cast and crew based on the performance of multiple movies in which the lead actors, actresses, and crew participated over a short-term (2 years before the release of the current movie), medium-term (5 years before the release of the current movie), and long-term (10 years before the release of the current movie). The same study used the director as the only key crew member, whereas our study also considers the producer, writer, and music composer in addition to the director as part of the key crew members.

2.4. Impact of lead cast and crew

Multiple empirical studies suggested two divergent views regarding the impact of lead actors on movie success. Some studies [44] found a significantly positive effect of stars on

movie performance. Other studies [45] failed to find any significant effects of the star cast on movie revenue. Few studies [46] established the negative effect of stars on movie revenue. Wallace et al. [47] established that some but not all the stars equally impacted movie revenue. Ravid [12] found that the influence of stars was insignificant on return-on-investments from movies. Our study analyses the impact of lead actors and crew members on AOVR.

3. Methodology

3.1. Data source and data processing

IMDb dataset is a popular database that has been used in multiple academic studies [2,9,41]. The dataset contains movie titles, regions, and languages. It also contains the details of the primary cast (actors, actresses) and crew (producer, writer, composer, director) members. It provides other details like the year of release, and multiple genres (maximum three). The dataset also assimilates the average rating of each movie on a scale of one to 10 where one is poor and 10 is the best rating. It provides the volume of ratings for each movie. We downloaded the IMDb dataset (https://datasets.IMDbws.com/) on 29th August 2020. We applied the following filters to churn out relevant data for our analysis.

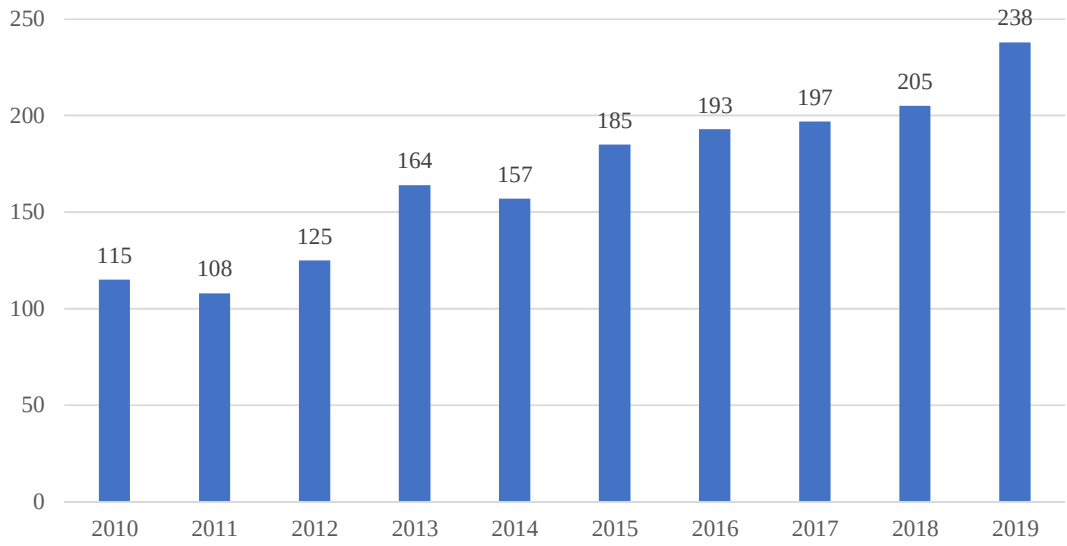
- Filter out titleType=="movie" from "title.basics" dataset
- Filter out region=="IN" from "title.akas" dataset. Region “IN” represents movies released in India
- Merge these above two datasets using the unique alphanumeric title identifier
- Include movies receiving at least 500 ratings.
- Remove non-Indian movies (Hollywood movies or foreign movies released in India)
- Consider movies released in India on or after 01st January 2010
- Exclude movies not having Runtime value
- Movies belonging to “Animation”, and “Documentary” genres are removed because professional actors are not involved [2]

The final dataset contains 1687 movies. Figure-1 shows the distribution by release year.

3.2. Data preparation

The primary objective of our study is to predict the AOVR of a movie much before its release. One of the primary predictors is the past performance of the key cast and crew members. We break down the key cast and crew members’ past performance as a combination of 1) performance in the recent past (last 2 years excluding the release year) 2) performance in the medium past (last 5 years excluding the release year), and 3) performance in the long past (last 10 years excluding the release year). The performance score of an individual lead cast or crew i (actor, actress, composer, director, producer, and writer) for year j is determined based on average rating and total volume of ratings received by all movies released in the previous n years, where that individual member participated. The following formulas depict the average rating for cast or crew, i for year, j:

Figure 1: Number of Movies by Release Year



Average Rating for long past$_{i,j}$ = $\sum_{j-10}^{j-1}\sum_{p=1}^{n} Rp,q,i$ / no of movies released for cast or crew i between year j-10 and j-1

Average Rating for medium past$_{i,j}$ = $\sum_{j-5}^{j-1}\sum_{p=1}^{n} Rp,q,i$ / no of movies released for cast or crew i between year j-5 and j-1

Average Rating Score for recent past$_{i,j}$ = $\sum_{j-2}^{j-1}\sum_{p=1}^{n} Rp,q,i$ / no of movies released for cast or crew i between year j-2 and j-1

where,

$R_{p,q,i}$ = Average rating of the pth movie released in year q, where the cast or crew i participated

The average rating for cast or crew i is converted to 0 in case there is no movie released during the period considered above.

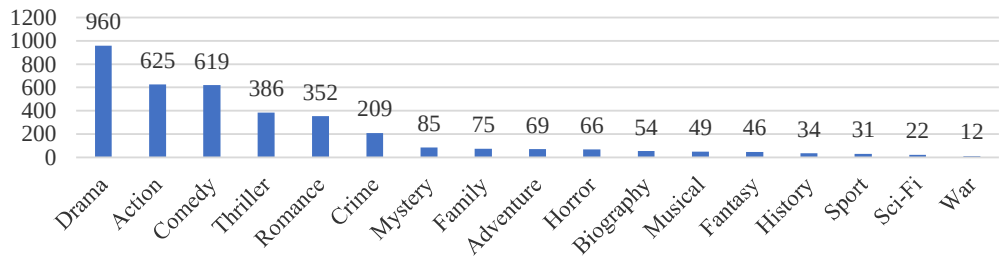
A movie employs one or multiple actors, actresses, composers, directors, producers, and writers, who are listed in the IMDb database. We introduce the following predictors to represent the long-term, medium-term, and short-term influence of the lead cast and crew in the movies. Detailed descriptions of the predictors are given in Appendix-A. Summary statistics of the lead cast and crew are given in Table-1 below.

Table 1: Summary Statistics of the lead cast and crew

Predictors	Mean	Median	Std Dev	Kurtosis	Skewness	Minimum	Maximum
actor_NT_Rating	4.42	4.91	2.09	-0.42	-0.73	0.00	8.20
actress_NT_Rating	3.74	4.53	2.73	-1.44	-0.29	0.00	8.60
composer_NT_Rating	3.35	4.15	3.00	-1.76	-0.06	0.00	8.50
director_NT_Rating	2.68	0.00	3.22	-1.59	0.48	0.00	8.90
producer_NT_Rating	2.64	0.00	2.90	-1.52	0.42	0.00	8.30
writer_NT_Rating	1.74	0.00	2.50	-0.31	1.09	0.00	8.60
actor_MT_Rating	4.75	5.42	1.98	0.28	-1.07	0.00	7.88
actress_MT_Rating	4.17	5.30	2.64	-1.15	-0.61	0.00	8.50
composer_MT_Rating	3.66	5.20	2.97	-1.70	-0.27	0.00	8.50
director_MT_Rating	3.92	5.30	3.21	-1.69	-0.25	0.00	8.85
producer_MT_Rating	3.18	3.31	2.95	-1.69	0.09	0.00	8.30
writer_MT_Rating	2.27	0.00	2.73	-1.19	0.65	0.00	8.60
actor_LT_Rating	4.85	5.54	1.92	0.60	-1.18	0.00	7.72
actress_LT_Rating	4.30	5.43	2.59	-0.99	-0.72	0.00	8.50
composer_LT_Rating	3.72	5.37	2.97	-1.69	-0.31	0.00	8.50
director_LT_Rating	4.28	5.63	3.10	-1.48	-0.49	0.00	8.85
producer_LT_Rating	3.32	3.65	2.94	-1.70	-0.01	0.00	8.30
writer_LT_Rating	2.45	0.00	2.79	-1.40	0.51	0.00	8.60

Another important aspect of a movie is the genre, which signifies the storyline. Our study picks all genres available on IMDb for each movie. A movie can be part of multiple genres. IMDb captures a maximum of three genres for each movie. The distribution of movies by genre type is given below in Figure-2. We include the top six genres for analysis. The rest of the genres are classified into other categories. Seven independent variables are introduced in our analysis representing each of the top six genres (Drama, Action, Comedy, Thriller, Romance, Crime) and the rest of the genres. The "Genre_Oth" variable combines the remaining genres outside the top six genres. A movie is coded one if the movie belongs to the specific genre and zero otherwise.

Figure 2: Number of Movies by Genre Type



We also include movie runtime as one of the predictors. Summary statistics of the predictor are given below in Table-2:

Table 2: Summary Statistics of Variable "Runtime":

Predictors	Mean	Median	Std Dev	Kurtosis	Skewness	Minimum	Maximum
Runtime	136.85	138	20.34	3.78	0.28	76	321

3.3. Data analysis

We assimilate 1687 movies released in India between 2010 and 2019. The cross-sectional dataset is randomly distributed into training datasets containing 75% of movies and test datasets containing the remaining 25% of movies. Kim et al [9] also partitioned data randomly into training and test datasets. Hyperparameters of each ML model are tuned using the training dataset, whereas model performance is evaluated using the test dataset.

We utilize multiple linear regression (MLR) as the baseline model to predict AOVR. The regression model determines the value of the dependent variable (Y) based on multiple independent variables (Xj) using the following equation.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_n X_n + e.$$

The values of β_0, β_1, β_2, ……, β_n are determined in such a manner that the sum of the squared error (e) is minimized. The significance of the relationship of each independent variable independently with the dependent variable is determined using a t-test and the associated p-value with the null hypothesis as $\beta_j = 0$. We have used the training dataset to determine the β_j values. The prediction error (absolute difference between the actual value and the predicted value) of the model is determined on the test dataset. Our study employs Root Mean Square Error (RMSE) as the prediction error matrix, which was also used in multiple other studies [2,9]. MLR also identifies significant predictors of AOVR.

We deploy Ridge and Lasso regression to reduce the error of the MLR method. The Ridge regression shrinks the value of the coefficients to reduce the model complexity and multi-collinearity [2]. This method adds a penalty to the sum of the squared error obtained in MLR as provided in the equation below:

Cost function for Ridge regression = sum of the squared error + $\lambda * (\beta_0^2 + \beta_1^2 + \beta_2^2 + \ldots.. + \beta_n^2)$

The penalty term (λ) regularizes the coefficients to optimize the cost function. It is also to be noticed that as $\lambda \rightarrow 0$, the Ridge regression becomes MLR. Similarly, as $\lambda \rightarrow \infty$, the coefficients (β) $\rightarrow 0$. It means that as lambda increase, variance decreases at the expense of bias. Therefore, it is important to optimize the value of lambda using the training dataset. We perform a 10-fold cross-validation method to optimize the lambda value. The optimum lambda value obtained during parameter tuning is 0.0722 for AOVR. Lower values suggest that the Ridge regression does not reduce prediction error in comparison to the MLR.

Lasso regression adds a penalty to the cost function as the sum of absolute coefficient values to the sum of squared error. Unlike Ridge regression, the Lasso regression method performs feature selection by forcing the coefficients of some of the predictors to zero [48]. Like in the Ridge regression, we perform a 10-fold cross-validation method to optimize the lambda value for the Lasso regression method. The optimum lambda value obtained during parameter tuning is 0.0053 for AOVR. Lower values suggest that the Lasso regression does not reduce prediction error in comparison to the MLR.

Elastic net is a combination of Lasso and Ridge regression, which add both L1 (Lasso) and L2 (Ridge) regularization as a penalty to the cost function. Like in Ridge and Lasso regression, we optimize the lambda value for both L1 and L2 regularization using a 10-fold cross-validation method. The optimum lambda values for AOVR are 0.4200 and 0.0155 for L1 and L2 regularization, respectively.

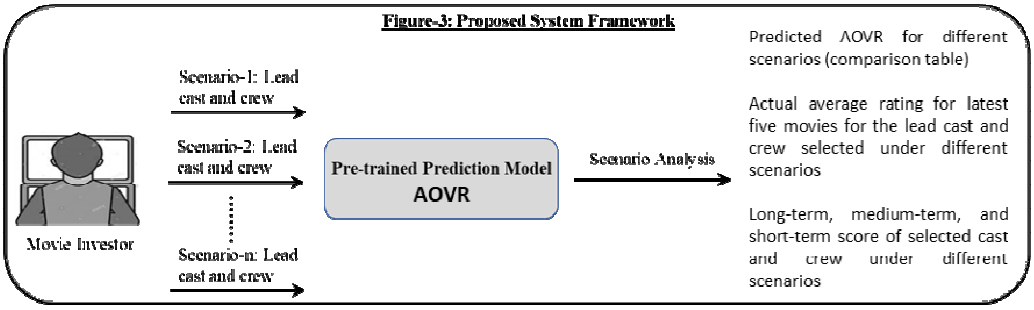
We introduce Support Vector Regression (SVR) as a supervised method to predict AOVR. SVR is a popular supervised ML algorithm, which was applied across multiple studies [2,9,49,50]. SVR creates a line or a hyperplane to fit the data within a pre-determined error margin. called the maximum error, ϵ (epsilon). Cost is another hyperparameter, which requires tuning in SVR to minimize over-fitting or under-fitting. SVR accepts more observations having errors higher than the epsilon value, as the cost increases and vice versa. We perform a grid search and a 10-fold cross-validation method to optimize both hyperparameters. Our study selects the combination of these two hyperparameters that generates the lowest RMSE value from 110 (11 different values for ϵ starting from 0.0 to 1.0 both inclusive with an increase of 0.1 and 10 different values for cost, which is represented as 2n, where n varies from 1 to 10, both inclusive with an increment of 1) different combinations for AOVR. The optimum combination is epsilon = 0.5, and cost = 2. We use the Radial kernel function with the Gamma value being kept constant at 0.0385 during the grid search.

Artificial Neural Network (ANN) is one of the most used ML algorithms in academics [30]. Input passes through multiple layers of connected neurons. The output of one layer becomes the input of the next layer. The weights of each neuron are calculated using the training dataset. We deploy both linear and non-linear (tanh, softmax, and sigmoid) activation functions. More hidden layers tend to overfit. Therefore, we restrict a maximum of three hidden layers each having between two and five neurons. The combination that has the lowest RMSE in the test dataset is selected as the final parameter for the ANN model.

In addition to the above ML algorithms, we use Random Forest (RF) as one of the ensembling methods. RF builds the trees independently of each other. We tune hyperparameters using a grid search method using 10-fold cross-validation.

3.4. Proposed solution framework

We propose a solution, where investors provide inputs (cast and crew members along with genre and runtime) for different scenarios. The pre-trained prediction model (updated periodically) computes AOVR for different scenarios, which are presented in a single comparison table. The comparison table assists the investors to identify the best scenario out of all possible alternatives. The solution also delineates the relative importance of different predictors of AOVR to improve model explainability and hence faster adoption of our proposed solution. The solution assimilates the actual AOVR of the latest 5 movies of the lead cast and crew for improved traceability. The proposed framework is depicted in Figure-3. The output of the proposed solution to be used by the movie investors for decision-making is provided in Appendix B.



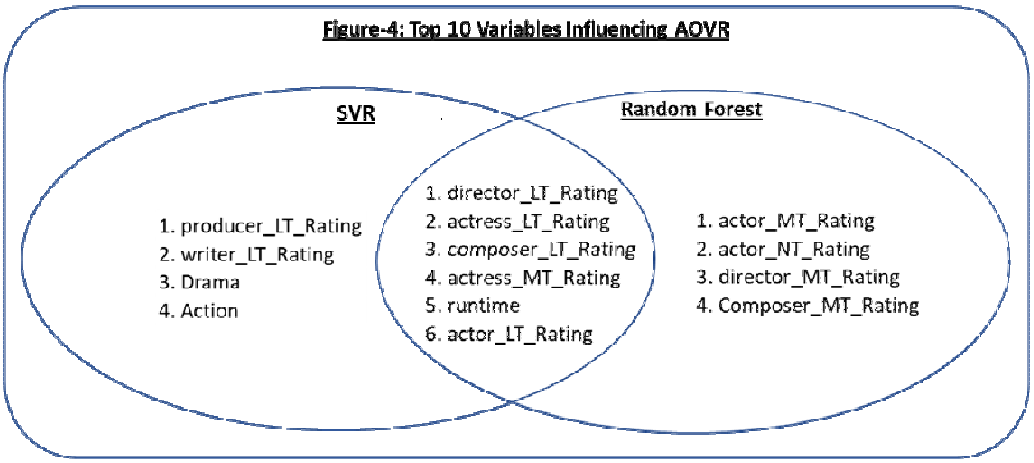
4. Findings

Our study includes multiple ML models to predict AOVR using short-term, medium-term, and long-term past performances of the movies of the key actors, actresses, and crew members. Table-3 below summarizes the error value (RMSE) of all the ML algorithms on the test dataset.

Table 3: Prediction Error Summary (RMSE) on test data by ML Algorithms

Sl No	Algorithms	RMSE
1	MLR (baseline model)	1.3672
2	Ridge	1.3169
3	Lasso	1.3166
4	Elastic Net	1.3149
5	SVR	1.2514
6	ANN	1.3268
7	RF	1.2276

The data clearly shows that RF is the best model for predicting AOVR. We utilize the two best models to identify the important parameters affecting AOVR, which is given in Figure-4.



The best ML model reduces the prediction error (RMSE) by 10.21% in comparison to the baseline model. Our study finds the long-term performance of the key actors, actresses, and crew as the primary predictors (refer to figure 4) of AOVR. Movie runtime and "drama", and "action" genres also fall into the top 10 influencers affecting AOVR as identified by one of the best two ML models. The two best models for AOVR find Long term performance of the composer as one of the top 10 influencers.

We provide a list of key influencers those impact movie performance, which increases the transparency of our solution and hence expedites the adoption of our solution. The prediction model developed by our study assists movie investors to predict viewers' ratings at the very early stage of movie production. This will help the investor to take more informed decisions by evaluating multiple scenarios based on the selection of the right cast and crew. This decision-making is extremely important as the fees of the key cast and crew form a significant chunk of investment.

5. Discussions

Our study improves the prediction accuracy of movie quality as represented by AOVR over the baseline model. Higher forecast accuracy improves the decision-making ability of the investors [51]. Identification of the influencing predictors gives additional insight to the investor to focus on a few key areas depending on the final objective. As an example, the selection of the music composer is important for scoring a higher average rating, justifying the importance of music and dance in Indian movies [17]. An investor needs to appreciate the inherent characteristics of movies from a specific country. The importance of music and dance cannot be established in the studies focussing on movies produced in Hollywood, and therefore, music composers are not important crew members for movie production. Our study for the first time assesses the importance of music composers on movie performance. We recommend investors give higher importance to the long-term performance of the key cast and crew than to the medium-term and short-term performance for better movie performance. To the best of our knowledge, there is a dearth of research in rating prediction for Indian movies. India, being the largest producer of movies is likely to see the adoption of our approach to predict movie performance at an early stage of movie production. The transparent reasoning of selecting one scenario over multiple other scenarios expedites the adoption of our proposed solution. Our solution also provides explainability to movie investors by highlighting the relative importance of cast and crew for better movie performance. Providing explainability enhances investors' trust in the predicted AOVR score. The fast adoption of online collaboration platforms like IMDb in India is likely to accelerate the usage of movie ratings as another key indicator of movie performance in addition to financial performance indicators like revenue. Our analysis supports the investors to explore this new platform (IMDb) more productively to reduce the risk of movie investment. The prediction system developed here can be applied across any Indian movie, be it in Hindi or other regional languages. There is an opportunity to customize the model specific to the language of the movie, like Hindi, Tamil, Bengali, or other regional languages. The customization of the prediction model is expected to improve prediction accuracy further. This same prediction approach can also be replicated in other mass media entertainment like TV series, web series, radio programs, music programs, and so on.

Acknowledgement. The author would like to thank the reviewers for their valuable comments on this paper.

REFERENCES

1. A. De Vany, and W. Walls, Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar, *Journal of Economic Dynamics & Control*. 28 (2004), 1035–1057–2
2. M.T. Lash and K. Zhao, Early Predictions of Movie Success: The Who, What, and When of Profitability, *Journal of Management Information Systems*, 33(3) (2016) 874–903, https://doi.org/10.1080/07421222.2016.1243969
3. S.M. Brewer, and J.M. Kelley and J.J. Jozefowicz, A blueprint for success in the US film industry, *Applied Economics*, 41(5) (2009) 589–606.
4. D.B. Jun, D.S. Kim and J.H. Kim, A Bayesian DYMIMIC model for forecasting movie viewers. *KAIST business school, working paper series* (KCB-WP-2011-003).
5. L. Angela, L. Yong and T. Mazumdar, Star power in the eye of the beholder: A study of the influence of stars in the movie industry, *Marketing Letters*, 25(4) (2014) 385-396
6. S.G. Dastidar and C. Elliott, The Indian film industry in a changing international market, *Journal of Cultural Economics*, 44 (2020) 97–116. https://doi.org/10.1007/s10824-019-09351-6
7. C. Elliott, P. Konara, H. Ling, C. Wang and Y. Wei, Behind film performance in China's changing institutional context: The impact of signals. *Asia Pacific Journal of Management*, 35(1) (2018) 63–95.
8. H.T. Keh, W. Ji, X. Wang, S. Joseph, and R. Singh, Online movie ratings: a cross-cultural, emerging Asian markets perspective, *International Marketing Review*, 32(3/4) (2015) 366-388, https://doi.org/10.1108/IMR-08-2013-0161
9. T. Kim, J. Hong and P. Kang, Box office forecasting using machine learning algorithms based on SNS data, *International Journal of Forecasting*, 31 (2015) 364–390
10. A. Elberse and J. Eliashberg, Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures, *Marketing Science* 22(3) (2003) 329–354.
11. D. Lovallo, C. Clarke and C. Camerer, Robust analogizing and the outside view: two empirical tests of case-based decision making, *Strategic Management Journal*, 33(5) (2012) 496–512.
12. S.A. Ravid, Information, blockbusters, and stars: a study of the film industry. *Journal of Business*, 72(4) (1999) 463—492.
13. K. Wen and C. Yang, Determinants of the box office performance of motion picture in China – indication for Chinese motion picture market by adapting determinants of the box office (part II), *Journal of Science and Innovation*, 1(4) (2011) 17–26.
14. ET Bureau, Film industry in India to hit $3.7 billion by 2020, (2017), Retrieved July 23 2020. https://economictimes.indiatimes.com/industry/media/entertainment/media/film-industry-in-india-to-hit-3-7-billion-by-2020-says-report/articleshow/60998458.cms
15. N. Mccarthy, Bollywood: India's Film Industry By The Numbers [Infographic], (2014), Retrieved August 12 2020.

https://www.forbes.com/sites/niallmccarthy/2014/09/03/bollywood-indias-film-industry-by-the-numbers-infographic/#7e1120e42488
16. Statista Report, Leading film markets worldwide in 2017, by number of tickets sold (in millions), (2018), Retrieved September 11, 2020. https://www.statista.com/statistics/252729/leading-film-markets-worldwide-by-number-of-tickets-sold/
17. S. Rao, I need an Indian touch: Glocalization and Bollywood films, *Journal of International and Intercultural Communication*, 3(1) (2010) 1-19. DOI: 10.1080/17513050903428117
18. L. Srinivas, The active audience: spectatorship, social relations and the experience of cinema in India, *Media, Culture & Society*, (2002). https://doi.org/10.1177/016344370202400201
19. G. Jones, N. Arora, S. Mishra and A. Lefort, Can Bollywood go global? *Harvard Business School discussion paper*, 9 (2008) 806-040.
20. Livemint, Steep star fees a challenge to south Indian cinema?, (2016) Retrieved September 19 2020. https://www.livemint.com/Consumer/l2sYtQ8WxJg4UrD7QG4t3O/Steep-star-fees-a-challenge-to-south-Indian-cinema.html
21. Livemint, Rs100 crore and counting: 10 years. 64 films, (2018), Retrieved September 19 2020, https://www.livemint.com/Consumer/beDJhiRjmTfMiyT1NNty5H/Rs100-crore-and-counting-10-years-64-films.html
22. Prabhakar, Business of Rs 100-cr films: Who gets what and why, *ET Bureau*, (2012), Retrieved September 07 2020. https://economictimes.indiatimes.com/industry/media/entertainment/business-of-rs-100-cr-films-who-gets-what-and-why/articleshow/15700710.cms
23. V. Polonski, Humans don't trust AI predictions - Here's how to fix it, (2018), Retrieved May 24 2019, https://www.oecd-forum.org/users/80891-dr-vyacheslav-polonski/posts/29988-humansdon-t-trust-artificial-intelligence-predictions-here-s-how-to-fix-it
24. R. Zhao, I. Benbasat and H. Cavusoglu, Do users always want to know more? Investigating the relationship between system transparency and users' trust in advice-giving systems, In *ECIS 2019 Proceedings*, Stockholm & Uppsala, Sweden, (2019) 1–12
25. P.K. Chintagunta, S. Gopinath and S. Venkataraman, The effects of online user reviews on movie box office performance: accounting for sequential rollout and aggregation across local markets, *Marketing Science*, 29(5) (2010) 944–957.
26. E. Rogers, New product adoption and diffusion, *Journal of Consumer Research*, 2(2) (1976) 290–301.
27. K. Jedidi, R.E. Krider C.B. Weinberg, Clustering at the movies, *Marketing Letters*, 9(4) (1998) 393–405.
28. C. Dellarocas, X.M. Zhang and N.F. Awad, Exploring the value of online product reviews in forecasting sales: the case study of motion pictures, *Journal of Interactive Marketing*, 21(4) (2007) 23–45.

29. Y. Lee, S-H. Kim and K.C. Cha, A generalized Bass model for predicting the sales patterns of motion pictures having seasonality and herd behaviour, *Journal of Global Scholars of Marketing Science: Bridging Asia and the World*, 22(4) (2012) 310–326.
30. D. Delen, R. Sharda and P. Kumar, Movie forecast guru: a web based DSS for Hollywood managers, *Decision Support Systems*, 43(4) (2007) 1151–1170.
31. L. Zhang, J. Luo and S. Yang, Forecasting box office revenue of movies with BP neural network, *Expert Systems with Applications*, 36(3) (2009) 6580–6587
32. K.J. Lee and W. Chang, Bayesian belief network for box-office performance: a case study on Korean movies, *Expert Systems with Applications*, 36(1) (2009) 280–291
33. F. Abel, E. Diaz-Aviles, N. Henze, D. Krause and P. Siehndel, Analyzing the blogosphere for predicting the success of music and movie products. 2010 *International Conference on Advances in Social Networks Analysis and Mining, Odense*, (2010) 276-280, doi: 10.1109/ASONAM.2010.50.
34. S.Asur and B.A.Huberman, Predicting the future with social media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, ON 2010*, (2010) 492-499, doi: 10.1109/WI-IAT.2010.63.
35. L. Qin, Word-of-blog for movies: a predictor and an outcome of box office revenue, *Journal of Electronic Commerce Research*, 12(3) (2011) 187–198.
36. Y. Chen and X. Jinhong, Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix, *Management Science*, 54(3) (2008) 477-91
37. M. Trusov, E.B. Randolph and P. Koen, Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking, *Journal of Marketing*, 73(5) (2008) 90-102
38. S. Moon, K.B. Paul and D. Iacobucci, Dynamic Effects among Movie Ratings, Movie Revenues, and Viewer Satisfaction, *Journal of Marketing*, 74(1) (Jan. 2010) 108-121
39. C-L. Liu, W-H. Hsaio, C-H. Lee, G-C. Lu and E. Jou, Movie rating and review summarization in mobile environment, *IEEE Trans. Syst. Man. Cybern.Part C (Appl. Rev.)*, 42(3) (2012) 397–407
40. W. Schmit and S. Wubben, Predicting ratings for new movie releases from twitter content, in: 6th Workshop on Computational Approaches to Subjectivity, *Sentiment and Social Media Analysis Wassa*, (2015) 122-126
41. A. Oghina, M. Breuss, M. Tsagkias and M. de Rijke, Predicting IMDB Movie Ratings Using Social Media, *In: Baeza-Yates R. et al. (eds) Advances in Information Retrieval. ECIR 2012*, Lecture Notes in Computer Science, 7224 (2012). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-28997-2_51
42. X. Ning, L. Yac, X. Wang, B. Benatallah, M. Dong and S. Zhang, Rating prediction via generative convolutional neural networks based regression, *Pattern Recognition Letters*, 132 (2020) 12-20. https://doi.org/10.1016/j.patrec.2018.07.028
43. P. Thomas, Are 'Baahubali' and '2.0' really expensive films or is it just marketing strategy?, (2016), Retrieved May 29 2020, http://www.newindianexpress.com/business/2016/nov/07/are-baahubali-and-20-really-expensive-films-or-is-it-just-marketing-strategy-1535769.html
44. A.L. Hadida, Commercial success and artistic recognition of motion picture projects, *Journal of Cultural Economics*, 34(1) (2010) 45–80.

45. S. Basuroy and S. Chatterjee, Fast and frequent: Investigating box office revenues of motion picture sequels, *Journal of Business Research*, 61 (2008) 798–803.
46. M. Fetscherin, The main determinants of Bollywood movie box office sales, *Journal of Global Marketing*, 23(5) (2010) 461–476
47. W.T. Wallace, A. Seigerman and M.B. Holbrook, The role of actors and actresses in the success of films: how much is a movie star worth? *Journal of Cultural Economics*, 17(1) (1993) 1-27
48. N. Ludwig, S. Feuerriegel and D. Neumann, Putting Big Data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests, *Journal of Decision Systems*, 24(1) (2015) 19-36. DOI: 10.1080/12460125.2015.994290
49. A. Ferencek and M.K. Borštnar, Data quality assessment in product failure prediction models, *Journal of Decision Systems*, 29(1) (2020) 1-8. DOI: 10.1080/12460125.2020.1776927
50. Y. Xiaopeng and D. Stanko, Data envelopment analysis may obfuscate corporate financial data: using support vector machine and data envelopment analysis to predict corporate failure for nonmanufacturing firms, *INFOR: Information Systems and Operational Research*, 55(4) (2017) 295-311, DOI: 10.1080/03155986.2017.1282290
51. A. Martínez, C. Schmuck, S. Pereverzyev, C. Pirker and M. Haltmeier, A machine learning framework for customer purchase prediction in the non-contractual setting, *European Journal of Operational Research*, 281(3) (2018) 588–596. https://doi.org/10.1016/j.ejor.2018.04.034.

APPENDIX

Appendix A: Description of predictors representing lead cast and crew

Sl No	Predictor Name	Description
1	actor_NT_Vote	Sum of volume of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), where the lead actors of the current movie also played the lead actor role.
2	actress_NT_Vote	Sum of volume of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), where the lead actresses of the current movie also played the lead actress role.
3	composer_NT_Vote	Sum of volume of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), where the lead composers of the current movie also composed songs.
4	director_NT_Vote	Sum of volume of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), which were also directed by the lead directors of the current movie
5	producer_NT_Vo	Sum of volume of ratings received by all the movies released

	te	between year t-2 and t-1 (where t is the release year of the current movie), which were also produced by the lead producers of the current movie
6	writer_NT_Vote	Sum of volume of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), which were also written by the lead writers of the current movie
7	actor_MT_Vote	Sum of volume of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), where the lead actors of the current movie also played the lead actor role.
8	actress_MT_Vote	Sum of volume of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), where the lead actresses of the current movie also played the lead actress role.
9	composer_MT_Vote	Sum of volume of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), where the lead composers of the current movie also composed songs.
10	director_MT_Vote	Sum of volume of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), which were also directed by the lead directors of the current movie
11	producer_MT_Vote	Sum of volume of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), which were also produced by the lead producers of the current movie
12	writer_MT_Vote	Sum of volume of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), which were also written by the lead writers of the current movie
13	actor_LT_Vote	Sum of volume of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), where the lead actors of the current movie also played the lead actor role.
14	actress_LT_Vote	Sum of volume of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), where the lead actresses of the current movie also played the lead actress role.
15	composer_LT_Vote	Sum of volume of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), where the lead composers of the current movie also composed songs.
16	director_LT_Vote	Sum of volume of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), which were also directed by the lead directors

		of the current movie
17	producer_LT_Vote	Sum of volume of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), which were also produced by the lead producers of the current movie
18	writer_LT_Vote	Sum of volume of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), which were also written by the lead writers of the current movie
19	actor_NT_Rating	Average of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), where the lead actors of the current movie also played the lead actor role.
20	actress_NT_Rating	Average of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), where the lead actresses of the current movie also played the lead actress role.
21	composer_NT_Rating	Average of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), where the lead composers of the current movie also composed songs.
22	director_NT_Rating	Average of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), which were also directed by the lead directors of the current movie
23	producer_NT_Rating	Average of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), which were also produced by the lead producers of the current movie
24	writer_NT_Rating	Average of ratings received by all the movies released between year t-2 and t-1 (where t is the release year of the current movie), which were also written by the lead writers of the current movie
25	actor_MT_Rating	Average of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), where the lead actors of the current movie also played the lead actor role.
26	actress_MT_Rating	Average of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), where the lead actresses of the current movie also played the lead actress role.
27	composer_MT_Rating	Average of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), where the lead composers of the current movie also composed songs.

28	director_MT_Rating	Average of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), which were also directed by the lead directors of the current movie
29	producer_MT_Rating	Average of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), which were also produced by the lead producers of the current movie
30	writer_MT_Rating	Average of ratings received by all the movies released between year t-5 and t-1 (where t is the release year of the current movie), which were also written by the lead writers of the current movie
31	actor_LT_Rating	Average of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), where the lead actors of the current movie also played the lead actor role
32	actress_LT_Rating	Average of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), where the lead actresses of the current movie also played the lead actress role.
33	composer_LT_Rating	Average of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), where the lead composers of the current movie also composed songs.
34	director_LT_Rating	Average of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), which were also directed by the lead directors of the current movie
35	producer_LT_Rating	Average of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), which were also produced by the lead producers of the current movie
36	writer_LT_Rating	Average of ratings received by all the movies released between year t-10 and t-1 (where t is the release year of the current movie), which were also written by the lead writers of the current movie

Appendix B: Single view of scenario analysis for the movie investors

Predicted AOVR and VOVR Across Multiple Scenarios

Prediction	Scenario-1	Scenario-2	Scenario-3	Scenario-4	Scenario-5	Scenario-6	Scenario-7	Scenario-8	Scenario-9	Scenario-10
AOVR										
VOVR										

Actual Performance of Last 5 Movies

Scenario-n ↓	Predicted AOVR	Predicted VOVR
Actor-1		
Actor-2		
Actress-1		
Actress-2		
Producer-1		
Director-1		
Composer-1		
Composer-2		
Writer-1		

Rating and Vote Score for Cast and Crew for Scenario-n

Scenario-n ↓	Short-term Rating	Medium-term Rating	Long-term Rating	Short-term Vote	Medium-term Vote	Long-term Vote
Actor-1						
Actor-2						
Actress-1						
Actress-2						
Producer-1						
Director-1						
Composer-1						
Composer-2						
Writer-1						