Movie Success Prediction using k-Nearest Neighbor Algorithm

Bramesh S M, Assistant Professor, P. E. S. College of Engineering, Mandya, India, brahmesh06s@gmail.com

Puttaswamy B S, Assistant Professor, P. E. S. College of Engineering, Mandya, India, bsputtaswamy2012@gmail.com

Abstract - **The quantity of movies produced and / or released in the globe is growing exponentially except during an outbreak of a pandemic disease like COVID-19. The success proportion of the movie is of highest importance since huge amount of money is financed in the creation of each of these movies. In such a situation, prior knowledge about the failure and / or success of a specific movie and knowing the factors affecting the movie success will benefit the production communities, since these forecasts will give them a reasonable idea of how to go about with the campaigning and advertising, which itself is an expensive issue altogether. So, the prediction of the failure and or success of a movie is very much important to the film industry. Hence, we aim to explore Internet Movie Database (IMDb) in order to predict the imdb_score, which will reflect the movie success, using k-Nearest Neighbor (k-NN) algorithm. From the experimental analysis, we found that k-NN on the test data got the highest accuracy of 68.64% when k = 9. Obtained results finally show that the prediction model using k-NN has achieved a decent prediction accuracy on the IMDb dataset.**

Keywords* — *k-NN, Predictive analytics, Movie success, Critical review, IMDb score, R.

I. INTRODUCTION

The global Box office revenue during the year 2019 was 42.2 billion USD [1] with many new movies produced every year. In such a scenario, there exist both successful movies and also failure movies. Consequently, how can we select a successful movie to relax and enjoy on our weekends? Usually, what we do to answer this question is? We look at the score of a movie or see its review on the website. The www.imdb.com website is just a good selection to refer at this time. The www.imdb.com contains information about movies and also the observations from viewers. The scores given on the IMDb website are extremely recognized by the public, reflecting the content excellence as well as the viewer's choice to some level. If one can come up with a model, which will predict the score on IMDb website, then this model would be a powerful model to identify how successful a movie will be? Even before it is released. Hence this research work, try to display the vital factors inducing the score on IMDb website and proposes an effective approach to forecast the movie success using the IMDb dataset [2].

II. LITERATURE SURVEY

Movie success mainly be subjected to the perceptions that, in what way the movie has been justified by the viewers. In the early period, many people ranked gross box office revenue ([3], [4], [5]). Few earlier work ([5], [6], [7]) has used IMDb data, to signify gross of a movie depending on stochastic and regression models. Few of them have categorized movies has either flop or success based on their profits and applied binary classifications for prediction. The extent of movie success does not exclusively depend on profits. Success of movies depends on a several issues like director, actresses / actors, background story, the time of release, etc. Further few researchers have developed a prediction model with some pre-released data which were used as their features [8]. In majority of the work, researchers have considered a very limited feature. Thus, their model's performance was poor. Moreover, they have also ignored involvement of viewers on whom success of a movie typically depends. Few researchers have also adopted several applications of Natural Language Processing (NLP) for sentiment analysis ([9], [10]) and collected movie reviews for their test domain. In such a scenario, the prediction accuracy lies on how large the test domain is. Having a domain of smaller size is not a better idea for measurement. Yet again, most of them did not consider critic's reviews into account. Also, viewers' reviews can be biased as a fan of actress / actor may fail to give an unbiased view. In [11] A. Sivasantoshreddy et. al has used hype analysis to predict a movie box-office opening prediction. In some cases, movie success prediction was made through neural network analysis ([8]). Some

DOI : 10.35291/2454-9150.2020.0565

researchers have come up with the prediction model using social media, social network and hype analysis ([12], [13]) where they calculated positivity and the number of comments related to a particular movie. The majority of the work seems to be focused towards analysis of movie reviews or user-specific preferences. However, using machine learning techniques to analyze the attributes of a movie is relatively unexplored method for predicting its success. Hence this paper aims at using IMDb 5000 movie dataset for analyzing the attributes of a movie and then to develop a prediction model to predict movie success using k-NN.

III. DATA DESCRIPTION

The IMDb dataset used in this paper is obtained from Kaggle website. The IMDb dataset has 28 variables for 5043 movies, spanning across a hundred years in sixty-six countries. In this work, we have used “imdb_score” has the output variable and other 27 variables have been used as possible input variables.

IV. PROPOSED METHODOLOGY

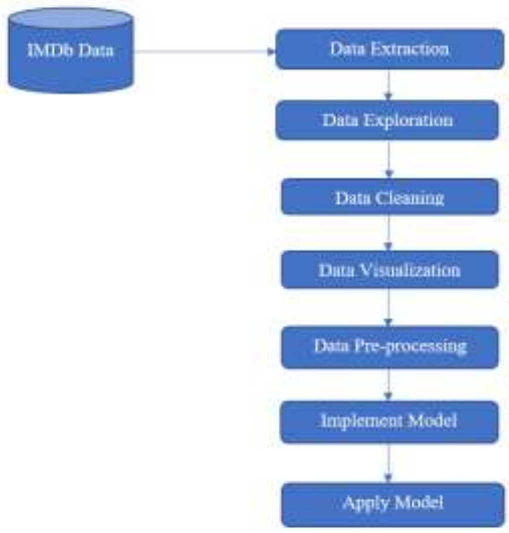

Fig. 1: The proposed methodology

As shown in Fig. 1, the proposed methodology includes the following steps:

A. Data Extraction

In this step, data is extracted (IMDb data) from Kaggle website in order to analyze which type of movies are more successful, in other words, get higher imdb_score?

B. Data Exploration

In this step, the data extracted in the previous step (IMDb data) is loaded into the R workspace by loading the suitable packages in R. After loading the IMDb data, we applied str(object, …) function in R on the IMDb data to understand the data at the high level. i.e., knowing about number of observations, number of variables in the IMDb data. We found that the IMDb data had 5043 observations and 28 variables. In this step, we also found that the response variable “imdb_score” was numerical, and the predictors were mixed with numerical and categorical variables. It is also observed that IMDb data had duplicates and hence we removed duplicates from the IMDb data. After removing the duplicates, we found number of observations left = 4998. We also observed that movie titles had (Â) as a special character at the end and some movie titles had whitespaces, hence we deleted them from the movie titles. Finally, in this step, we found that there was a small amount of difference in the averages of imdb_score related to diverse genres, and also found that nearly all the averages are in the same range of 6 ~ 8. So, we removed input variable “genres” as it was not really related to the imdb_score.

C. Data Cleaning

In this step, firstly we dealt with missing values in each column, by using colSums() function in R to total NA in every column. Then we used a heatmap to visualize the missing values.

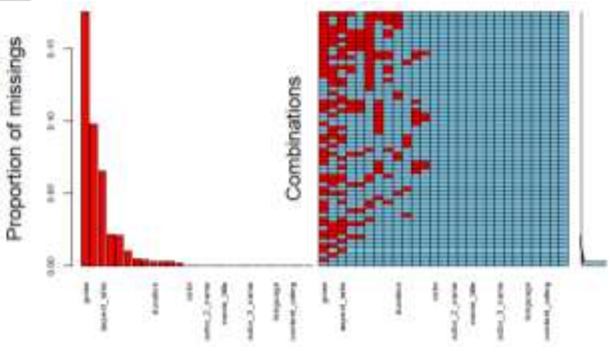

Fig. 2: Heatmap to visualize missing values

From the heatmap, we found that the budget and gross input variables have many missing values, and as we want to retain these two variables for further analysis, we deleted rows with null values for budget and gross. The result of this operation is not too bad, because this operation omitted 23% of the observations. Now our dataset has 3857 observations. Secondly, we analyzed aspect_ratio input variable and found that aspect_ratio input variable had the highest number of missing values in the IMDb dataset. Then from the means of imdb_score for diverse aspect_ratios, we found that there was no significant difference, and all the means was falling in the range of 6.3 ~ 6.8. Hence, we removed this input variable as it won’t affect our following analysis. Then, we dealt with 0s and NAs. i.e., in the case of numeric predictors, we deal with 0s and NAs by replacing 0s with NA, and then replacing all NAs with their respective column mean and in the case of categorical predictors, Blanks (“”) represent missing values and these missing values cannot be replaced with sensible data, hence we deleted those rows. Since IMDb data has gross and budget information, we added two columns: profit and percentage return on investment for further analysis. We removed input variables like colored and language from further analysis as it was observed that both input variables were almost constant. Finally, around 8% movies are from UK, 79% from USA, 13% from other countries, we clustered other countries together to make this categorical input variable (countries) with less levels: UK, USA, Others.

D. *Data Visualization*

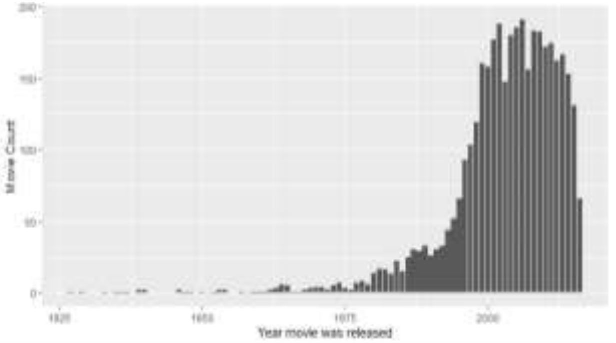

Fig. 3: Histogram of Movie Released

As we can see from Fig. 3 that production of movies just exploded after year 1990. It might be due to improvement in technology and commercialization of the internet.

From the Fig. 3, we can also see that there are not many records of movies released before 1980. Hence, we removed those records because they might not be representative in this case.

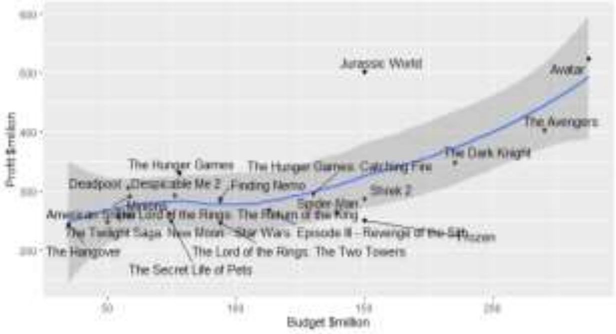

Fig. 4: **Top 20 movies based on its Profit**

Fig. 4 shows the top twenty movies based on the Profit earned (Gross - Budget). From Fig. 4 it can also be observed that highly budgeted movies tend to earn higher profit. i.e., the trend is nearly linear in this case.

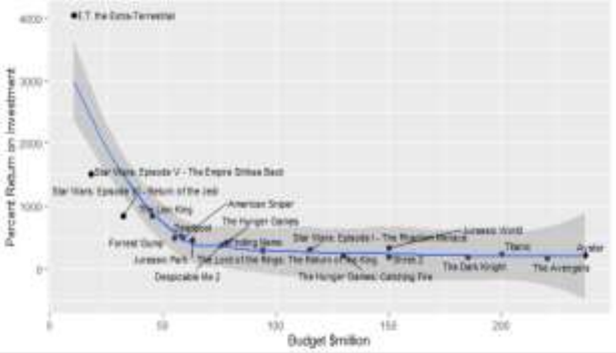

Fig. 5: Top 20 movies based on its Return on Investment

Fig.5 shows the top twenty movies based on its Percentage Return on Investment. Since movie profit does not give a clear picture about its financial success over the years, this analysis, over the absolute value of the Return on Investment (ROI) across its Budget, would provide good results. Analysis on the Commercial Success acclaimed by the movie (Gross earnings and profit earned) versus its imdb_score was also performed. From the analysis, we found that, there was not much correlation between them.

E. *Data Pre-processing*

Since the IMDb dataset has 1660 directors, 3621 actors and all the names are so different for the whole dataset, we removed the names from further analysis. Same with plot_keywords input variable, i.e., from the analysis we found that plot plot_keywords are too diverse to be used in the prediction and also found that the movie link was a redundant variable. Hence, we removed these two variables from further analysis.

In order to avoid multicollinearity, here we removed profit and return_on_investment_perc variables from further analysis. Then we removed Highly Correlated Variables with the help of correlation heatmap. Since our goal was to build a model, which can help us to predict if a movie is good or bad. Means we don't really want an exact imdb_score to be predicted, we only want to know how good or how bad is the movie. Therefore, we bin the imdb_score into four buckets: less than 4, 4 ~ 6, 6 ~ 8 and 8 ~ 10, which represents bad, OK, good and excellent respectively. Finally, we prepared IMDb data for applying k-NN purpose. i.e., dummy variables are used for categorical variables after that we normalized IMDb data. Here we have split IMDb dataset into training, validation and test sets with the ratio of 6:2:2.

F. *Implement k-Nearest Neighbor algorithm*

To implement a prediction model, we have used k-Nearest Neighbour classification method from a package called Fast Nearest Neighbor search algorithms and applications (FNN:) in R.

The distance metric used is the Euclidean distance. The model predicts the success of a movie based on the majority vote concept, with ties broken at random.

G. *Apply Model*

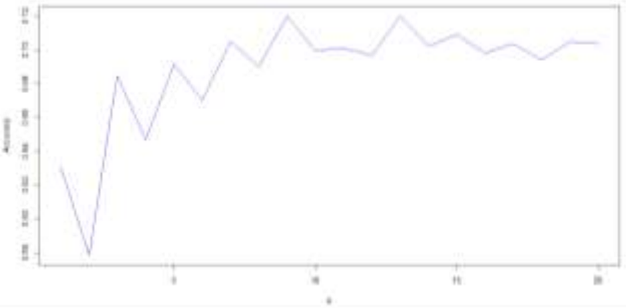

Fig. 6: Accuracy of k-NN for different k-value based on Euclidean distance on normalized validation data

DOI : 10.35291/2454-9150.2020.0565

As shown in the Fig. 6, to find the best k value, we varied k value from 1 to 20, and build twenty different prediction models on validation data and calculated each model's prediction accuracy on validation data. Finally, we found that the prediction model got the highest accuracy of 71.96%, when k = 9 and k = 13. Since the prediction model took less time when k = 9, when compared with the k =13, we selected the best value of k as 9. After this we applied prediction model on the test data by setting k = 9. In this case the prediction model gives the accuracy of 68.64%.

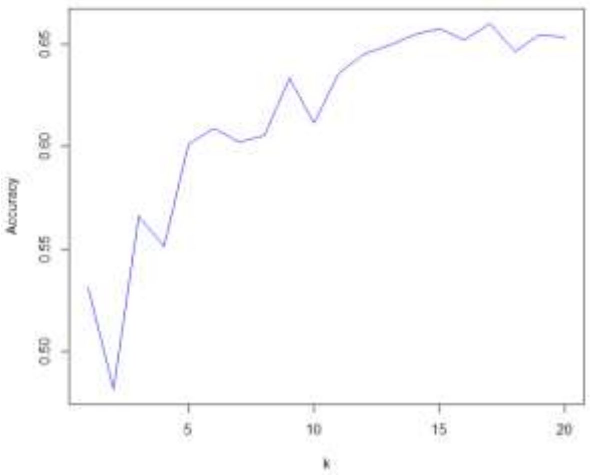

Fig. 7: Accuracy of k-NN for different k-value based on Euclidean distance on raw validation data

As shown in the Fig. 7, again to find the best k value, we varied k value from 1 to 20, and build twenty different prediction models on raw validation data (i.e., without using the normalization technique) in this case and calculated each model's prediction accuracy on raw validation data. Finally, we found that the prediction model got the highest accuracy of 66.03%, when k = 17. This investigation shows that normalizing the IMDb data and then applying the k-NN model provides better accuracy when compared with applying the k-NN model without normalizing the IMDb data.

V. Conclusion

A successful movie entertains viewers, and also allows film firms to gain greater profits. A portion of factors such as experienced actors, good directors is considerable for creating good movies. However, famous actors and good directors can continuously bring an expected box-office revenue, but cannot guarantee a highly rated imdb_score. The IMDb dataset is a motivating dataset to predict the imdb_score, which will reflect the success of a movie. Based on the obtained results, we can say that the k-NN algorithm accuracy depends on the k-value and it also depends on whether we are applying k-NN on normalized data or applying k-NN without normalizing the data. That means k-value cannot be chosen randomly, instead it should be calculated carefully by doing an investigation. Also, through this investigation we found that the user vote, Facebook likes are important factors that play key roles in predicting success of a movie. In conclusion, high user vote, high Facebook likes is positively correlated to higher imdb_score, which reflects success of a movie. In the future, we try to increase the features count and also movies count in the dataset. We would also try to comprise other sources of movie data collection, such as YouTube and Twitter.

References

[1] https://www.statista.com/statistics/259987/global-box-office-revenue/#statisticContainer

[2] https://data.world/data-society/imdb-5000-movie-dataset

[3] S. Gopinath, P. K. Chintagunta, and S. Venkataraman, "Blogs, Advertising, and Local-Market Movie Box Office Performance," Management Science, vol. 59, no. 12, pp. 2635–2654, 2013.

[4] M. C. A. Mestyán, T. Yasseri, and J. Kertész, "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data," PLoS ONE, vol. 8, no. 8, 2013.

[5] J. S. Simonoff and I. R. Sparrow, "Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers," Chance, vol. 13, no. 3, pp. 15–24, 2000.

[6] A. Chen, "Forecasting gross revenues at the movie box office," Working paper, University of Washington, Seattle, WA, June, 2002.

[7] M. S. Sawhney and J. Eliashberg, "A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures," Marketing Science, vol. 15, no. 2, pp. 113–131, 1996.

[8] R. Sharda and E. Meany, "Forecasting gate receipts using neural network and rough sets," in Proceedings of the International DSI Conference, pp. 1–5, 2000.

[9] B. Pang and L. Lee, "Thumbs up? Sentiment classification using machine learning techniques," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, pp. 79–86, July 2002.

[10] P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches," in Proceedings of the Hawaii International Conference on System Sciences (HICSS), 2005.

[11] A. Sivasantoshreddy, P. Kasat, and A. Jain, "Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining," International Journal of Computer Applications, vol. 56, no. 1, pp. 1–5, 2012.

[12] J. Duan, X. Ding, and T. Liu, "A Gaussian Copula Regression Model for Movie Box-office Revenue Prediction with Social Media," Communications in Computer and Information Science Social Media Processing, pp. 28–37, 2015.

[13] L. Doshi, J. Krauss, S. Nann, and P. Gloor, "Predicting Movie Prices Through Dynamic Social Network Analysis," Procedia - Social and Behavioral Sciences, vol. 2, no. 4, pp. 6423–6433, 2010.

DOI : 10.35291/2454-9150.2020.0565