Movie Success Prediction using Machine Learning Algorithms and their Comparison

1st Rijul Dhir
Department of Computer Science
IIT Patna
Bihar, India
rijul.dhir@gmail.com

2nd Anand Raj
Department of Computer Science
NIT Jalandhar
Punjab, India
anand.v.raj95@gmail.com

***Abstract*—The number of movies produced in the world is growing at an exponential rate and success rate of movie is of utmost importance since billions of dollars are invested in the making of each of these movies. In such a scenario, prior knowledge about the success or failure of a particular movie and what factor affect the movie success will benefit the production houses since these predictions will give them a fair idea of how to go about with the advertising and campaigning, which itself is an expensive affair altogether. So, the prediction of the success of a movie is very essential to the film industry. In this proposed research, we give our detailed analysis of the Internet Movie Database (IMDb) and predict the IMDb score. This database contains categorical and numerical information such as IMDb score, director, gross, budget and so on and so forth. This research proposes a way to predict how successful a movie will be prior to its arrival at the box office instead of listening to critics and others on whether a movie will be successful or not. The proposed research provides a quite efficient approach to predict IMDb score on IMDb Movie Dataset. We will try to unveil the important factors influencing the score of IMDb Movie Data. We have used different algorithms in the research work for analysis but among all Random forest gave the best prediction accuracy which is better in comparison to the previous studies. In the exploratory analysis we found that number of voted users, number of critics for reviews, number of Facebook likes, duration of the movie and gross collection of movie affect the IMDb score strongly. Drama and Biopic movies are best in genres.**

Keywords— Box office gross; Data Mining; Machine learning; Movie success; Movie; Predictive analytics; Critical review; Rating.

I. INTRODUCTION

Nowadays, hundreds of movies are produced and released every year. Among them, there exist both great movies and crappy ones. Therefore, how do we know their qualities before we do not see the movie ourselves? Or how can we choose a great movie to enjoy and relax on our weekends? Most of the time, we will turn to the movie score or have a look at its review to decide. IMDb website is just a good choice to refer at this time. Due to its popularity, IMDb website contains a great deal of information about movies and the comments from audiences. The scores which IMDb gives are highly recognized by the public, representing the quality of content as well as audience's favor to some extent. Therefore, in this research, we will try to unveil the important factors influencing the score on IMDb website and propose an efficient approach to predict it. The data we use in our paper comes from IMDb 5000 Movie Dataset on Kaggle. It contains 28 variables for 5042 movies and 4906 posters, spanning across 100 years in 66 countries. There are 2399 unique director names and thousands of actors/actresses. The worldwide Box office revenue 2016 was 38.3 billion USD [1] with hundreds of new movies made each year. If one can use a computer to predict how successful a movie will be even before it is released, this would be a powerful tool to use. An aspiring Hollywood director or a movie studio with some technical skills could predict whether their movie idea is going to be a safe investment. With the vast amount of data published on the Internet and the increasing power of the modern computer, is it possible to take advantage of these resources to make predictions.

The study can be used as a proof of concept for applications in other areas, and should highlight some of the challenges one needs to overcome to successfully create a prediction model. This idea could in theory be extended to predict credit ratings, the stock market or housing market. The only requirement being a vast and reliable data source. When combining the questions mentioned above to form a problem statement, formulating good as a measurement of a movies rating and sales, the following problem statement was produced. Movie ratings in recent years are influenced by many factors that makes the accurate prediction of ratings for the new movies being released a difficult task. There also have been various semantic analysis techniques to analyze user reviews which were applied to analyze the IMDb movie ratings. None of the studies has succeeded in suggesting a model good enough to be used in the industry. In this project, we attempt to use the IMDb dataset to predict the Cinema has a profound impact on our society. Cinema is one of the most powerful media for mass communication in the world. Cinema has the capacity to influence society both locally and globally. Many different kinds of movies are made every year. Some movies portray historical events, some create a culture, while some provide fantasy, and some do many more.

We perform an exploratory analysis of the data and observe some interesting phenomenon, which also helps us improve our prediction strategy as well as we will get to know about the features which affect the movie IMDb score. Our results finally show that we achieve a good prediction accuracy of IMDb score on this dataset

II. LITERATURE SURVEY

Success of a movie primarily depends on the perspectives that how the movie has been justified. In early days, a number of people prioritized gross box office revenue ([2], [3], [4]), initially. Few previous work ([4], [5], [6]), portend gross of a movie depending on stochastic and regression models by using IMDb data. Some of them categorized either success or flop based on their

revenues and apply binary classifications for forecast. The measurement of success of a movie does not solely depend on revenue. Success of movies rely on a numerous issue like actors/actresses, director, time of release, background story etc. Further few people had made a prediction model with some pre-released data which were used as their features [7]. In most of the case, people considered a very few features. As a result, their models work poorly. However, they ignored participation of audiences on whom success of a movie mostly depends. Although few people adopt many applications of NLP for sentiment analysis ([8], [9]) and gathered movie reviews for their test domain. But the accuracy of prediction lies on how big the test domain is. A small domain is not a good idea for measurement. Again most of them did not take critics reviews in account. Besides, users' reviews can be biased as a fan of actor/actress may fail to give unbiased opinion. M. T. Lash and K. Zhao's [10] main contribution was, firstly they developed a decision support system using machine learning, text mining and social network analysis to predict movie profitability not revenue. Their research features several features such as dynamic network features, plot topic distributions means the match between "what" and "who" and the match between "what" and "when" and the use of profit based star power measures. They analyzed movie success in three categories, audience based, released based and movie based. Their hypothesis based on the more optimistic, positive, or excited the audiences are about a movie, the more likely it is to have a higher revenue. Similarly, a movie with more pessimistic and negative receptions from the public may attract fewer people to fill seats. They retrieve data from different types of media. Such as Twitter, comments from YouTube, blogs, new 5 articles and movie reviews, star rating from reviews, the sentiment of reviews or comments have been used as a means for assessing audience's excitement towards a movie. Their original dataset collected from both Box-office Mojo and IMDb. They focused on the movies released in USA and excluded all foreign movies from their experiment. In A neural network had been used in the prediction of financial success of a box office movie before releasing the movie in theaters. This forecasting had been converted into a classification problem categorized in 9 classes. The model was represented with very few features. In [11] A. Sivasantoshreddy, P. Kasat, and A. Jain tried to predict a movie box-office opening prediction using hype analysis.

A neural network had been used in the prediction of financial success of a box office movie before releasing the movie in theaters [12]. This forecasting had been converted into a classification problem categorized in 9 classes. The model was represented with very few features. In [13], it was tried to improve movie gross prediction through News analysis where quantitative news data generated by Lydia (high-speed text processing system for collecting and analyzing news data). It contained two different models (regression and k-nearest neighbor models). But they considered only high budget movies. The model failed if common word used as name and it could not predict if there were no news about a movie. M.H Latif, H. Afzal [14] who used IMDB database only as their main source and their data was not clean. Again their data was inconsistent and very noisy as mentioned by them. So they used Central Tendency as a standard for filling missing values for different attributes. K. Jonas, N. Stefan, S. Daniel, F. Kai use sentiment and social network analysis for prediction [15] their hypothesis was based on intensity and positivity analysis of IMDb sub forum Oscar Buzz. They had considered movie critics as the influencer and their predictive perspective. They used bag of word which gave wrong result when some words were used for negative means. There was no category award and only concerned with the award for best movie, director, actors/actress and supporting actors/actress. In some cases, success prediction of a movie was made through neural network analysis ([7], [18]). Some researchers made prediction based on social media, social network and hype analysis ([16], [17], [19], [20]) where they calculated positivity and number of comments related to a particular movie. Moreover, few people had predicted Box Office movies' success based on Twitter tweets and YouTube comments.

III. MOTIVATION

As data scientists we wanted to dig deeper into the business side of movies and explore the economics behind what makes a successful movie. Basically we wanted to examine whether there are any trends among films that lead them to become successful at the box office, and whether a film's box office success correlates with its ratings. A useful analysis would help us predict how well a film does at the box office before it screens, without having to rely on critics or our own instinct. Essentially we want to determine if there is a "Hollywood formula" to making a successful movie. How can we tell the greatness of a movie before it is released in cinema? This question puzzled me for a long time since there is no universal way to claim the goodness of movies. Many people rely on critics to gauge the quality of a film, while others use their instincts. But it takes the time to obtain a reasonable amount of critics' review after a movie is released. And human instinct sometimes is unreliable. Analyzing the attributes of a movie using machine learning techniques is a relatively unexplored method for predicting its success. Even considering that such information might be of interest not only to the movie sector in the form producers and financiers, but also to academics, service providers and viewers, most of the current work seems to be focused towards user-specific preferences or analysis of movie reviews.

IV. DATA DESCRIPTION

The dataset we utilized to train and test our model came from kaggle.com. The dataset includes information about

several movies on IMDb, including movie titles, directors, genres, countries of origin, and the Facebook popularity of the top three actors featured in the film. This information was provided in a variety of formats, including strings, integers, and floating point data. In order to implement the machine learning algorithms effectively and avoid underutilization of certain aspects of the movies provided in the dataset, the data was converted to numerical values using the Scikit learn preprocessing library to scale the features.

The dataset contains 28 variables for 5043 movies, spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses. “imdb_score” is the response variable while the other 27 variables are possible predictors. IMDb website is just a good choice to refer at this time. Due to its popularity, IMDb website contains a great deal of information about movies and the comments from audiences. The scores which IMDb gives are highly recognized by the public, representing the quality of content as well as audience’s favor to some extent. Roughly speaking, half of the variables is directly related to movies themselves, such as title, year, duration, etc. Another half is related to the people who involved in the production of the movies, e.g., director names, director face book popularity, movie rating from critics, etc.

V. PROPSED METHODOLOGY

The first step is to identify a dataset of movie data that’s representative and suitable for analysis. Relevant attributes of such data must include general pre-production information regarding film productions such as genre, language and information about the actors and directors involved. Likewise, the data must also include some measure of success, such as user originated movie ratings. Secondly, the relevant dataset has to be prepared and structured in such a way that the data used is representative of the movie scene at large, as well as viable for analysis by the relevant machine learning techniques and algorithms. Lastly, the prediction performance of the relevant machine learning algorithms has to be evaluated based on the specified dataset. This means that a set of suitable tools has to be acquired, as well as configured for evaluating both algorithms in comparison to each other based on the data, whilst still ensuring equivalence between in measurements. To be able to compare the algorithms based on their prediction performance, suitable measures of this parameter must therefore also be identified.

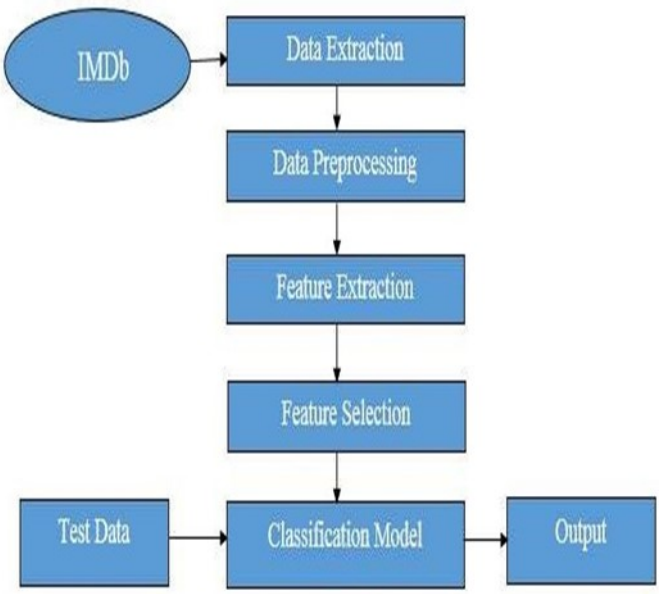

Figure 1. Workflow of research

A. *Correlation*

The correlation is one of the most common and most useful statistics. A correlation is a single number that describes the degree of relationship between two variables. Let's work through an example to show how this statistic is computed. There exist several different correlation techniques. The Survey System's optional Statistics Module includes the most common type, called the Pearson or product-moment correlation. The module also includes a variation on this type called partial correlation. The latter is useful when you want to look at the relationship between two variables while removing the effect of one or two other variables.

Like all statistical techniques, correlation is only appropriate for certain kinds of data. Correlation works for quantifiable data in which numbers are meaningful, usually quantities of some sort. It cannot be used for purely categorical data, such as gender, brands purchased, or favorite color. If we have positive correlation, then it means that the two variables are simultaneously moving in the common direction and if there is a negative correlation then the two variables are moving in inverse or opposite direction. The correlation matrix is used to find the relationship between all the variable with each other in the dataset.

VI. CLASSIFIER

A. *Support Vector Machine*

Support Vector Machine (SVM) is a supervised machine learning model that is used for classification. SVMs work by maximize the margin between separating hyper plane. In linear SVM the plane can be split by a line, see figure 3.14 for an example how the model could look like. For example, could the red values be answer A and the blue be answer B. If a new value would be introduced to the system and positioned on the red side, the model would predict the new value to be equal to answer A. If there are more answers possible a hyper plane is created to be able to split all the

answers up in different areas. SVM are effective high dimensional, memory efficient, and versatile machine learning algorithms that work well with non-linear data.

B. Random Forest

It is ensemble algorithm. Ensemble algorithms are those which combines more than one algorithms of same or different kind for classifying objects. For example, running prediction over Naive Bayes, SVM and Decision Tree and then taking vote for final consideration of class for test object. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. Irregular timberlands RF or arbitrary choice woodlands are a gathering learning strategy for characterization, relapse and different errands, that work by developing a huge number of choice trees at preparing time and yielding the class that is the method of the classes (in arrangement) or mean forecast (in relapse) of the individual trees. RF is an improvement over the decision tree algorithm as it corrects habit of over fitting in decision trees to their training set.

C. Ada Boost

Ada-boost, like Random Forest Classifier is another ensemble classifier. (Ensemble classifier are made up of multiple classifier algorithms and whose output is combined result of output of those classifier algorithms). Ada-boost classifier combines weak classifier algorithm to form strong classifier. A single algorithm may classify the objects poorly. But if we combine multiple classifiers with selection of training set at every iteration and assigning right amount of weight in final voting, we can have good accuracy score for overall classifier retrains the algorithm iteratively by choosing the training set based on accuracy of previous training. The weight-age of each trained classifier at any iteration depends on the accuracy achieved. Each weak classifier is trained using a random subset of overall training set.

D. Gradient Boost

Extreme Gradient Boosting (XG Boosting) [21] is one of the implementations of the Gradient Boosting, but there is something which makes it different from Gradient boosting is the control the overfitting by using the more regularized model which help in more accurate prediction. The name XG Boost though actually refers to the engineering goal to push the limit of the computation resources for boosted tree algorithm. Which is the reason why many of the people use XG Boost algorithm. For the model, it might be more suitable to be called as regularized gradient boosting.

E. K-Nearest Neighbors

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data.

VII. EXPERIMENTAL RESULT AND DISCUSSION

A. Correlation analysis

It is a statistical method to evaluate and study the strength of a relationship between two, numerically measured variables. It is done to check if there are possible connections between variables.

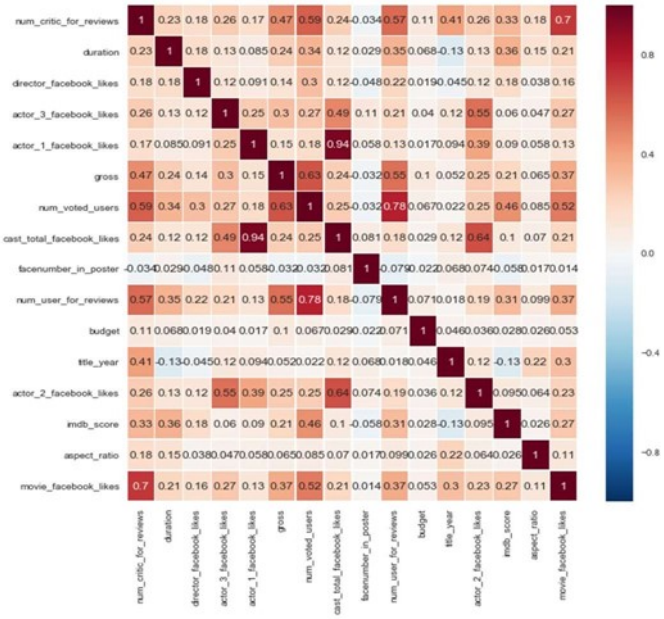

Figure 2. Heatmap of Correlation Matrix

Positive Correlation

- Number of Critic for reviews
- Duration
- Gross
- Number of voted Users
- Number of Facebook Likes
- Movie's Genres: Biography, Drama, Historical things Negative Correlation
- facenumber_in_poster title_year B

B. Classifier Results

Table 1 shows that Random forest gives the best accuracy among all other algorithms followed by gradient boost. It represents the detailed parameter values for the various algorithm which includes Precision, Recall, F-measure, and accuracy.

Table 1: Evaluation of Classifier

	Precision	**Recall**	**F1 Score**	**Accuracy**
Random Forest	61	60	59	61
Gradient Boost	58	57	56	56.68
KNN	46	44	45	44.3
SVM	50	46	45	45.88
Ada Boost	49	49	49	49.15

Figure 3 represents the accuracy comparison of different algorithms. From the above graph it is clear that random forest and gradient boost are the two algorithms which are giving the best accuracy. So after the analysis we came to know that random forest is predicting the movie success more accurately than any other algorithm on the following data set. Ada boost classifier, SVM and KNN were also used for the prediction but the overall result were not satisfactory. The accuracy of Ada boost was 49%, SVM was 45% whereas the lowest accuracy was of KNN classifier.

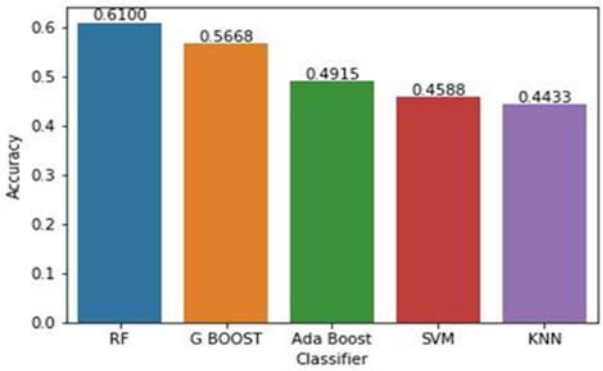

Figure 3. Accuracy comparison of classifier

Random forest gave the best results in comparison to the other algorithms as it takes random sample of data as an input and select the subset of features randomly. So it is robust in case of over fitting while boosting algorithms are sensitive to label noise as it fits classification model to an exponential loss. Random Forest algorithms are trained with random sample of data (even more randomized cases available like feature randomization) and it trusts randomization to have better Boosting algorithms are sensitive to label noise as it fits classification model to an exponential loss. Random Forests are trained with random sample of data (even more randomized cases available like feature randomization) and it trusts randomization to have better generalization performance on out of train set. It can discover very complex dependences and large data (over fit less likely with RF).

C. Result analysis with previous work

Our proposed model has given better accuracy in comparison to the previous studies done on predicting the movie success. Additional feature included to this study improved the performance of the system. In the past studies prediction made on Korean movies gave 58.5% [31] of accuracy on the other hand the accuracy on IMDb date between 2000 to 2012 gave only 50.7%[28] of accuracy. In our model after including the unexplored features and taking more number of date the system gave the best accuracy.

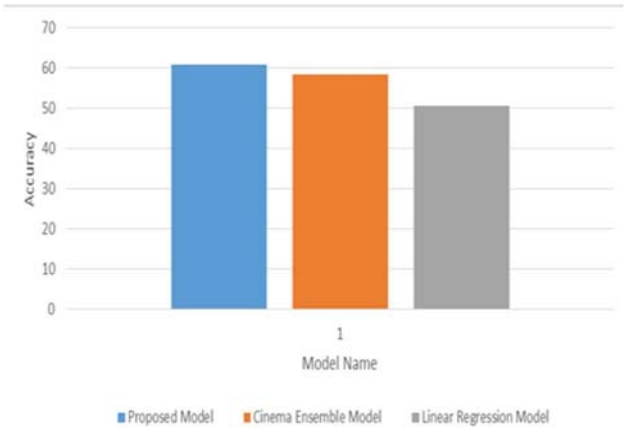

Figure 4. Accuracy comparison of classifier

From Figure 4 it is clear that our model is better in comparison to the previous model with respect to accuracy. The above chart shows the comparison of best accuracy achieved in the past research.

VIII. CONCLUSION AND FUTURE WORK

The IMDb dataset is an interesting dataset to analyze. After building the five models we found out that the Random Forest represents the movie features more accurately. The success percentage for all models are better in comparison to the previous studies. The results obtained are better than that of some standard libraries and similar studies. A movie success does not only depend on features related to movies. Number of audience plays very important role for a movie to become successful. Because the whole point is about audiences, the whole industry will make no sense if there is no audience to watch a movie. Number of ticket sold during a specific year can indicate the number of audiences of that year. In the future, we would like to increase both the number of movies and features in the dataset. We would also like to include other social media sources of movie data collection such as Twitter and YouTube. Other learning models that we want to apply to the movie data are the following supervised learning models: MLP and Bagging. We are interested in comparing results from these models with those expressed herein.

REFERENCES

[1] "Global box office revenue 2016 | Statistic." [Online]. Available: https://www.statista.com/statistics/259987/global-box-office-revenue/. [Accessed: 03-Jun-2018].

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] M. T. Lash and K. Zhao, "Early Predictions of Movie Success: The Who, What, and When of Profitability," Journal of Management Information Systems, vol. 33, no. 3, pp. 874–903, Feb. 2016.

[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] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," Expert Systems with Applications, vol. 30, no. 2, pp. 243–254, 2006.

[13] W. Zhang and S. Skiena, "Improving Movie Gross Prediction through News Analysis," 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009.

[14] M.H Latif, H. Afzal "Prediction of Movies Popularity Using Machine Learning Techniques", National University of Sciences and technology, H-12, ISB, vol. 16, no. 8, pp. 127–131, 2016.

[15] K. Jonas, N. Stefan, S. Daniel, F. Kai "Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis" University of Cologne, Pohligstrasse 1, Cologne, pp. 23-26,2013.

[16] 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.

[17] 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.