iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 10 **Issue:** V **Month of publication:** May 2022

**DOI:** https://doi.org/10.22214/ijraset.2022.42653

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 10 Issue V May 2022- Available at www.ijraset.com*

# IMDB Box Office Prediction Using Machine Learning Algorithms

Mohini Gore[5], Aishwarya Sheth[2], Samrudhi Abbad[3], Paryul Jain[4], Prof. Pooja Mishra[1]
[1, 2, 3, 4, 5]*Department of Computer Engineering*

***Abstract: Movies are a big part of our world! But nobody knows how a movie will perform at the box office. There are some bix budget movies that bomb and there are smaller movies that are smashing successes. This project tries to predict the overall worldwide box office revenue of movies using data such as the movie cast, crew, posters, plot keywords, budget, production companies, release dates, languages, and countries. The dataset on Kaggle contains all these data points that you can use to predict how a movie will fare at the box office. Among many movies that have been released, some generate high profit while the others do not. This paper studies the relationship between movie factors and its revenue and build prediction models. Besides analysis on aggregate data, we also divide data into groups using different methods and compare accuracy across these techniques as well as explore whether clustering techniques could help improve accuracy***
***Keywords: component: regression; predictive analytics; Clustering; Expectation-maximization; K-means; Movies***

## I. INTRODUCTION

The income of film industry comes from screening movie in the theater, which is called “Box-Office”. Film industry is a highly competitive industry. Many new movies queue up to be released each week, so a theater owner has to decide on which movie to be shown, based mainly on revenue. Regression analysis is a widely-used technique to predict revenue. This paper aims to compare accuracy among various types of regression analysis, with and without clustering techniques. Specifically, for regression analysis, we used three different types of regression to create prediction models, which are linear regression, polynomial regression, and support vector regression. As clustering into smaller groups of similar items may improve accuracy of the prediction models, we cluster the movie data and apply regression analysis on each cluster. Three clustering techniques are explored, namely clustering by movie genre, K-mean clustering and Expectation Maximization clustering. In order to compare performance across models, k-fold cross validation technique was employed to divide the data in two groups: training and testing data. The training data was used to create the prediction model, while the testing data was used to test accuracy of the regression models. R-square and root mean square error (RMSE) were used as performance indicators of the models.

## II. RELATED WORKS

There exist many research works attempting to predict movie revenue using variations on the regression models, such as different techniques or factors, in order to investigate which approach could generate the highest accuracy. To perform linear regression, many used classic factors such as cast, producer, director and genre, while some employed other additional different factors. For example, in [1], classic factors were used along with social media data such as Facebook or Twitter to predict movie revenue. In addition, [2] used additional factors, such as Motion Picture Association of America (MPAA) rating, the number of screens and holiday-released date, while [3] proposed a model using MPAA rating and criticism from audiences to build a regression model for movie with over 50,000 dollars in revenue. Moreover, a research in China showed that directors had more influence on movie revenue than actors by using multiple linear regression [4]. To improve accuracy, [5] showed that doing regression after clustering based on budget and the number of theaters that show the movie could decrease error. However, [6] found that linear regression might not always work well because the dataset could be too small to generate an accurate model. Therefore, [7] used polynomial regression along with classic factors to predict movie revenue and discovered that a higher degree in equation might lead to more error. Support vector regression (SVR), one of non-linear regression methods, was shown to provide higher accuracy when compared with linear regression, ridge regression and logistic regression .

In [2], clustering by Expectation Maximization (EM) method was used to divide data into groups based on the number of theaters that showed a movie in the first week. The regression method was then applied to predict revenue of movie in each group. This study found that applying EM clustering onto data before doing regression decreased the prediction error. To consider influence of movie stars and directors on revenue, [4] used the number of movie star appearances as a parameter for each movie star called “star power” and averaged the star powers between leading actor and actress for each movie. In [6], the same approach was applied to directors of the movies. Furthermore, the Oscar Award was used to represent the influence of the movie stars in [10].

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 10 Issue V May 2022- Available at www.ijraset.com*

## III. METHODOLOGY AND PROPOSED MODELS

This paper aims to study and compare various methods for movie revenue prediction. We develop the prediction models using its related factors by applying different kinds of regression analysis. Moreover, this study also explores the concept of grouping data using different techniques before performing regression to study whether the approach can improve prediction accuracy.

This section explains our proposed approach and models in four major steps. First, data source and preparation are described. Second, we start developing models by applying three types of regression, i.e. linear regression, polynomial regression, and SVR, onto the aggregate cleaned data. Third, we take the same dataset and divide data into groups using three methods, i.e. by movie genres, using EM clustering, and K-means clustering. In addition to EM clustering, which was found to produce good results [2], we have added grouping by genre and K-means clustering, a commonly used clustering technique, for further technique comparison. Finally, we examine the model performance using R-square and RMSE.

### *A. Data source and Preparation*

In this study, the data used for analysis comes from a public database at https://www.kaggle.com/tmdb/tmdb- movie-metadata/data. We first examined the data and eliminated attributes which are not useful for our analysis, such as the movie's homepage, the movie's id, etc. We separated 10 remaining attributes into two major types: numerical attributes and non- numerical attributes, as shown in Table I.

Table I. Movie Data Attributes

| Numerical attributes (5) |
|---|
| Budget, Revenue, Vote Average, Vote Count, Runtime |
| Non-numerical attributes (5) |
| Genres, Spoken Language, Production Companies, Release Date, Cast |

Our original dataset contains 4804 movies. We then removed the movie data with missing values because accuracy may decrease if we use incomplete data to perform regression. We also removed the movie data with total revenue lower than 100,000 dollars as they were hard to predict [3] and were insignificant in our case. After data cleansing, 3121 movies were remained for analysis.

Next, we considered both major types of attributes. Numerical attributes were ready for the analysis in the next step, but non-numerical attributes were not. Therefore, we converted these non-numerical attributes to numerical attributes to be used for regression. For Genres, Spoken Languages, and Production Company is Walt Disney Pictures, the data values were converted as shown in an example in Table II.

Table II. Example of Conversion From Non-Numerical to Numerical Attributes

| | Aladdin | Now You See Me 2 | Johnny English Reborn |
|---|---|---|---|
| Budget ($M) | 28 | 90 | 45 |
| Revenue ($M) | 504 | 335 | 160 |
| Vote Average | 7.4 | 6.7 | 6 |
| Vote Count | 3416 | 3235 | 1007 |
| Runtime (min) | 90 | 129 | 101 |
| Genres | | | |
| Adventure | 1 | 1 | 1 |
| Action | 0 | 1 | 1 |
| Romance | 1 | 0 | 0 |
| Spoken Language | | | |
| English | 1 | 1 | 1 |
| Chinese | 0 | 1 | 1 |
| Production Companies | | | |
| Walt Disney Pictures | 1 | 0 | 0 |
| Summit Entertainment | 0 | 1 | 0 |
| Universal Pictures | 0 | 0 | 1 |

From the Released Date attribute, we compute its day of the week, call it Day of Week attribute and use binary variables for each day of week in order to explore whether there exists a relationship between a movie's day of week of the release date and its revenue. For example, we convert 14 September 1995 to Thursday.

Table III. Example of Prepared Data

| | Aladdin | Now You See Me 2 | Johnny English Reborn |
|---|---|---|---|
| Budget($M) | 28 | 90 | 45 |
| Revenue ($M) | 504 | 335 | 160 |
| Vote Average | 7.4 | 6.7 | 6 |
| Vote Count | 3416 | 3235 | 1007 |
| Runtime(min) | 90 | 129 | 101 |
| Star Power | 21 | 16 | 6 |
| Genres | | | |
| Adventure | 1 | 1 | 1 |
| Action | 0 | 1 | 1 |
| Romance | 1 | 0 | 0 |
| Spoken Language | | | |
| English | 1 | 1 | 1 |
| Chinese | 0 | 1 | 1 |
| Production Companies | | | |
| Walt Disney Pictures | 1 | 0 | 0 |
| Summit Entertainment | 0 | 1 | 0 |
| Universal Pictures | 0 | 0 | 1 |
| Day of week | | | |
| Wed | 1 | 0 | 0 |
| Thu | 0 | 1 | 1 |

For the Cast attribute, we convert it to numerical attribute using star power. Our star power value of a movie is a sum of star power values of two leading cast members. Star power value of a cast member is calculated by counting the number of occurrences of the star in the dataset. For example Companies, we used binary variables for each value of these attributes. For example, Aladdin's genre is Adventure and Romance, its Spoken Language is English, and its Production, Aladdin's leading casts are Robin Williams and Scott Weinger. Robin Williams has star power value of 20 because there are 20 Robin Williams movies in this dataset, and Scott Weinger has star power of 1 by the same analogy. Therefore, the star power value of Aladdin is 21. An example of completely prepared data is shown in Table III.

*B. Aggregate Data Analysis*

After the dataset was fully prepared, it is ready for regression analysis. In an aggregate data analysis, different kinds of regression were applied to the entire movie data. Our study explores three types of regression analysis:

1) *Linear Regression:* The objective of linear regression is to find the response value by generating a linear equation as a function of predictor variables$\beta$i denote co-efficients of predictor variables $x$i , such as Budget, Vote Average, etc.
2) *Polynomial Regression:* This non-linear regression models relationship between dependent variable and independent variables that may be non-linear.

The three proposed grouping techniques are as follows:

*a)* *Grouping Movie by Their Genres*: There are 18 movie genres, including Action, Adventure, etc., in this dataset, so there are a total of 18 groups. A movie may belonged to several genres; therefore, a movie and its data may appeared in several groups. In our study, all numerical attributes, namely Budget, Revenue, Vote Average, Vote Count, and Runtime, were used together as the factors for clustering.

*b)* *Random Forest Regression:* Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. To get a better understanding of the Random Forest algorithm, let's walk through the steps: Pick at random k data points from the training set. Build a decision tree associated to these k data points. Choose the number N of trees you want to build and repeat steps 1 and 2. For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model. $R^2$ score tells us how well our model is fitted to the data by comparing it to the average line of the dependent variable. If the score is closer to 1, then it indicates that our model performs well versus if the score is farther from 1, then it indicates that our model does not perform so well

*c)* *Support Vector Regression (SVR):* This regression method helps alleviate an overfitting problem of the regression method and may increase accuracy of the model. Support vector regression selects the widest gap between data points and create a hyperplane to separate data into two groups

*d)* *Group-based Analysis:* Our study also investigates whether grouping data before applying the regression onto each group will yield a higher accuracy than applying the regression directly to all data together. To achieve this goal, we proposed three models to group data. Specifically, we proposed segmenting movie data by its genre, as revenue of movie type may be driven by different factors. Moreover, we proposed grouping data using two different clustering techniques, namely EM clustering and K-means clustering. EM clustering has been shown to produce a good prediction [2], while K-means clustering is a popular clustering tool in data mining. Before applying the two clustering techniques, the data was normalized in order to adjust values of different scales into a common scale. For example, Budget value could be more than a million whereas Vote Average value would be less than 10. Therefore, the data were normalized so that the clustering techniques can work across various dimensions or attributes

*e)* *Using K-means Clustering:* This technique is one of the most popular and commonly used technique for clustering. The algorithm is quite simple. The value of k was specified by trial and error, such as 2, 3, etc. Then, k observations were randomly selected as centroid points for each cluster. Each data point was assigned to the cluster whose centroid is nearest to that data point (minimum Euclidean distance). The new centroid was then calculated from the current members of each cluster. These processes were done iteratively until the total of Euclidean distance between data points and their clusters' centroid was minimized, and the clusters' centroids converge. This technique assigns data to clusters deterministically, contrary to the EM technique. Again, all numerical attributes were used as the factors for clustering.

## IV. PROJECT SCOPE

Data fetching, cleaning the data and converting it into structured data to be analyzed easily. Project will be successful in predicting accurate revenue and help theater owner prioritize movie releases. Future work may include the integration of social factor data, such as reviews on website, into the analysis to further improve the accuracy of the prediction.

## V. ADVANTAGES

It can perform both regression and classification tasks. A random forest produces good predictions that can be understood easily. It can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm. Ideal in the following situations:

*1)* Predict the movie revenue.
*2)* Effective prediction technique
*3)* Beneficial for accurate prediction of which movies has to be released on priority basis.
*4)* Secure and efficient system

## VI. DISADVANTAGES

A country's GDP rate can be used as a feature to know if there is financial stability during the period when a movie is released. During an economic depression, very few amount of audience will go to the theatres to enjoy movies. So these facts play a vital role in an ultimate success of a movie.So this can't be predicted. When using a random forest, more resources are required for computation. It consumes more time compared to a decision tree algorithm.

## VII. LIMITATIONS

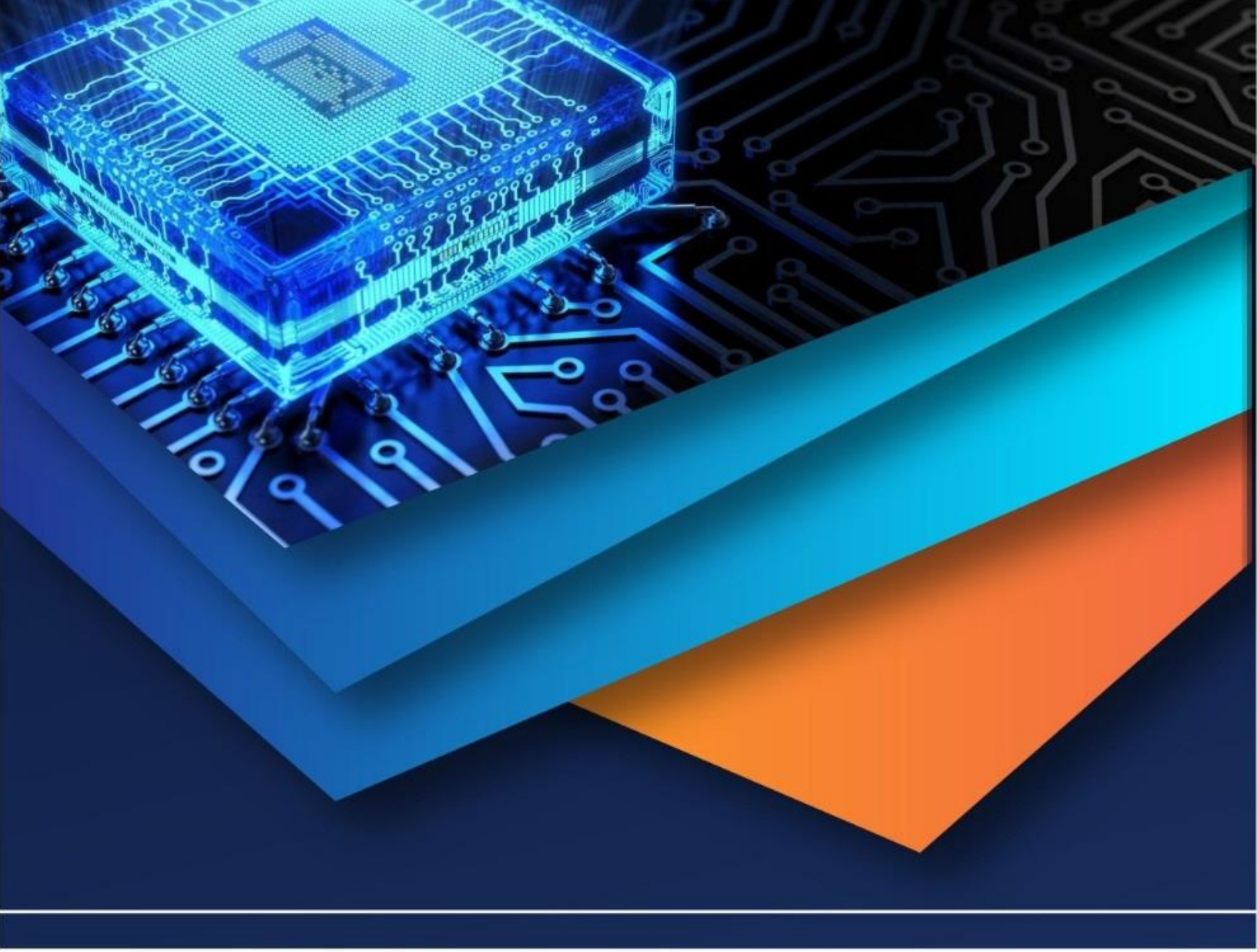
The number of audience plays a vital role for a movie to become successful. Because the whole point is about viewers, the entire industry will make no sense if there is no audience to watch a movie. The number of tickets sold during a specific year can indicate the number of viewers of that year. And the role of movie audience depends on many situations like political conditions and economic stability of a country. A country's GDP rate can be used as a feature to know if there is financial stability during the period when a movie is released. During an economic depression, very few amount of audience will go to the theatres to enjoy movies. Random forest does not produce good results when the data is very sparse. In this case, the subset of features and the bootstrapped sample will produce an invariant space. This will lead to unproductive splits, which will affect the outcome. So these facts play a vital role in an ultimate success of a movie. So in this situation our model doesn't predict the revenue

## VIII. CONCLUSION

Predicting movie revenue can be done by using various data analysis techniques. Regression analysis is one of the techniques that uses numerical data to predict the related response. In this paper, we apply various regression techniques onto movie data, with and without grouping, in order to predict the movie revenue. We found that the linear regression without applying clustering technique is the most accurate method in terms of R-square. Moreover, dividing movie data by its genre before doing regression can improve revenue prediction accuracy in some cases. However, applying clustering techniques, such as EM and K-means, can improve revenue prediction accuracy in terms of RMSE. We achieved an accuracy score of approximately 81%. Let's compare this to the scores we got with previous regression models: Simple Linear Regression: 50% Multiple Linear Regression: 65% Decision Tree Regression: 65% Support Vector Regression: 71% Random Forest Regression: 81% We can see that our Random Forest Regression model made the most accurate predictions thus far with an improvement of 10% from the last model.

## REFERENCES

[1] A. Bhave, H. Kulkarni, V. Biramane, and P. Kosamkar, "Role of different factors in predicting movie success," 2015 International Conference on Pervasive Computing (ICPC), pp. 1-4, June 2015.
[2] G. He and S. Lee, "Multi-model or Single Model? A Study of Movie Box-Office Revenue Prediction," 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, pp. 321-325, 2015.
[3] D. Im and M. T. Nguyen, "PREDICTING BOX-OFFICE SUCCESS OF MOVIES IN THE U.S. MARKET," CS229, Stanford University, Fall 2011
[4] Y. Yongbin and O. Rongzhao, "A study on the relationship among the leading actors, directors, and the box office income of a film — Based on multiple linear regression model," 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering, Xi'an, pp. 469-471, 2013.
[5] S. Shim and M. Pourhomayoun, "Predicting Movie Market Revenue Using Social Media Data," 2017 IEEE International Conference on Information Reuse and Integration (IRI), San Diego, CA, pp. 478-484, 2017.
[6] S. Yoo, R. Kanter, D. Cummings, and A. Maas, "Predicting Movie Revenue from IMDb Data," 2011.
[7] N. Apte, M. Forssell, and A. Sidhwa, "Predicting Movie Revenue," CS229, Stanford University, December 16, 2011.
[8] C. Lee and M. Jung, "PREDICTING MOVIE SUCCESS FROM SEARCH QUERY USING SUPPORT VECTOR REGRESSION METHOD, " International Journal of Artificial Intelligence & Applications (IJAIA)., Vol. 7, No. 1, January 2016.
[9] B. Flora, T. Lampo, and L. Yang, "Predicting Movie Revenue from Pre-Release Data," CS229, Stanford University, December 12, 2015
[10] K. Taewan, J. S. Uk, and D. Son, "Influence of Star Power on Movie Revenue. Global Journal of Emerging Trends in e-Business, Marketing and Consumer Psychology," 2016



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY