IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

Analyzing and Predicting The Success Of Box Office Collection Of A Movie Using Machine Learning

Sanyam Jatale[1], Rohan Moze[2], Varsha Khandekar[3],Shubham Jain[4] and Sanket Mokate[5]
Students, Department of Information Technology[1,2,3,4,5]
Smt. Kashibai Navale College of Engineering, Pune, India

Abstract: *Our day-to-day life has always been influenced by what people think. Ideas and opinions of others have always affected our own opinions. The explosion of Web 2.0 has led to increased activity in Podcasting, Blogging, Tagging, Contributing to RSS, Social Bookmarking, and Social Networking. The motion picture industry is a multi-billion-dollar business, and there is a massive amount of data related to movies are available over the internet. The framework will foresee an estimated achievement pace of a film dependent on its productivity by dissecting verifiable information from various sources like IMDB, Rotten Tomato, Box Office Mojo and Metacritic. Utilizing distinctive AI calculations, Machine Learning Tools, and different procedures the framework will foresee a film box office benefits depending on certain highlights like caste, genre, budget, actors, and many more features. 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. Thus, predicting the box-office will help this growing industry experts to imply some important business decisions in order to make the upcoming movie more successful.*

Keywords: Movie Industry, Machine Learning, Random Forest Classifier, XGBoost Regressor, Decision Tree, Prediction, Grid_Search, Cross-Validation, Recursive Feature Elimination, Director, Actor, Movie_title.

I. NOMENCLATURE

- ML – Machine Learning
- RF – Random Forest
- Sk – Scikit learn
- RFE – Recursive Feature Elimination
- XGB – XGBoost model
- DT – Decision Tree

II. INTRODUCTION

Movie revenue depends on various components such as cast acting in a movie, budget for the making of the movie, film critics review, rating for the movie, release year of the movie, etc. Because of these multiple components there is no formula that helps us to provide analysis for predicting how much revenue a particular movie will be generating. However, by analyzing the revenues generated by previous movies, a model can be built which can help us predict the expected revenue for a particular movie.[7] Such a prediction could be very useful for the movie studios which will be producing the movie so they can decide on different expenses like artist compensations, advertising of the movie,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

promotions in various cities, etc. accordingly. Plus, it allows investors to predict an expected return-on-investment (ROI). Also, it will be useful for many movie theatres to estimate the revenues they would generate from screening a particular movie Now a day's, online review system has become one of the most important part of any business approach. Posting reviews online for products bought or services received has become a trendy approach for people to express opinions and sentiments, which is essential for business intelligence, vendors, and other interested parties. Social media contains rich information about people's preferences.[8]

In this study, builds a prediction model for movie's success based on machine learning techniques, presents high-level of prediction accuracy. However, their efforts to improve the models' prediction power have been limited only to the modification of the algorithms rather than finding meaningful features that might be critical to expect the success of movie. To elaborate, the researchers in the past have mainly focused on introducing new machine learning algorithms and testing their performances, and it was pretty much the sole objective of their studies. Although such efforts have contributed to the increase of the prediction accuracy, we believe that the accuracy can be further increased by taking other perspectives. For example, it is possible to introduce an unexplored feature to a prediction model or to implement a feature-selection for existing features. We propose a decision support system for movie investment sector using machine learning techniques. In this project we will generate the predictive models based on machine learning tools and statistic analytical models. According to the calculated rate we will classify the movie into hit, average, or flop. Also using a regression model we'll predict the imdb_score for the movie.[2]

Through this project we aim to provide a machine learning algorithm which gives the most accurate result for movie success prediction.

III. EXISTING WORK

There are various papers that give ideas about analyzing and predicting the success of a movie's box office collection. The research shows that these articles and papers prove to be very useful for determining the improved versions and methodologies for better results and appropriate analysis for movie predictions. As we look all over these research papers, we get to know the diversities in approaches and methodologies. Also, it will help us in locating a roadmap for our project requirements and save us quite a useful time. People all over the world do some different work over the same project and can build something extraordinary which will help in solving some real-life complexities. From this existing model, we have come up with our approaches to make this project a great success.

From the absolute most recent examination papers like [8], [6], [11] we came to know the various methodologies with the end goal of the forecast. From here we can assemble data and thoughts dependent on various insights. Methods utilized are Sentiment Analysis, Prediction Using Data Mining, Machine learning models. These papers have assisted us with characterizing our venture's scope and aided in necessity investigation and prerequisite social occasions. A portion of the scientists included Neural Networks for forecasts which came out to be exceptionally precise yet there were a few issues we needed to confront. A counterfeit dataset should be made which was tedious interaction. After investigating sufficient papers, we came to discover subtleties of the task execution which facilitated our way.

During the implementation, we experienced various issues related to the procedures. In a portion of the papers where AI models were utilized, either expectation precision was not proper or highlight choice strategies were not right. Likewise, numerous ventures have less measure of information which was not a productive method to manage forecast models. Finding and investigating various procedures and models from different papers we wound up giving a successful and proficient model in our undertaking.

In [6] we have read about the Movie Success Prediction Using Data Mining where we came across a problem that using this we can only determine the relationship between different features. This paper helped us to reduce the number of parameters based on their dependencies. While reading [11] we have clearly understood the sentiment analysis and classification techniques but the main problem was the predictions are dependent on people's predictions and opinions.

As we do not have the best features, we got to know while reading SVM, Random Forest, and Regression used for models in [4] that for successful prediction we need to capture the best features. Authors in [9] proposed us the idea about the poor accuracy on using Predicting Movie Success Based on IMDB Data But as we have some methods for

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

improving accuracy, we have selected this method to use.

Paper [5] was about predicting movie ratings of IMDb users but the main problem was the lack of appropriate data for efficient accuracy. As we have decided to use IMDb historic data, [1] seeds ideas about the prediction movie rating of IMDB warned us about the predictions, which are also sometimes falsely made.We could have used the automatic movie rating prediction as cited in [10] that statistical functions are used and finalized the better resulting method. In this way, with this much existing work, we have made our decision to take methods and accuracy for getting excellent accuracy.

In a literature survey, we came to know that the techniques used by researchers are sentiment analysis, classification techniques, natural language processing, statistical modeling, support vector machine, random forest method, etc. But the problems related to some of their predictions depend on people's perception, predictions are for the only selected language, accuracy is less.

We will attempt to create a model that can anticipate more precisely an achievement of film industry expectations. For this, we utilize The Jupyter Notebook as an instrument and python language for coding reasons.

IV. PROPOSED METHODOLOGIES

Generally, the feature selection is one of the frequently-considered methods to increase the performance and the interpretability of machine learning algorithms. However, in this study, we focus more on introducing a new 3 feature rather than pruning the expectation model with existing ones. The reasoning behind our decision is that the features used in our study already have been tested to be highly effective for predicting a movie success in the past research. Also keeping in mind, the features, we select should be appropriate and suitable as per the building the predicting model. Thus, we expect that the exclusion of some of such features will decrease the accuracy of the prediction model. In addition, we have considered that the number of features used in this study is not as many to the extent that it deteriorates the performance of a prediction model. Elimination of the features will be done by analyzing factors between the features by carrying out some graphs and plots (bar plot, boxplot, pair plot etc.). Using confusion matrix and such tools to determine best possible features for the model. For prediction we'll be completely relying to the machine learning algorithms such as Random Forest Classifier, decision Tree, Regression model and the predicted results will be validated using Log loss, Grid Search etc. At last, determining the best model and using it for the important aspect that is for prediction of box-office success of a new movie prior to its release.

As of now there are many methodologies designed in order to determine the prediction of a movie. Our methodologies include some machine learning tools and statistical analytics for business decision support system.

4.1 General Framework of Proposed System

Following Fig, I show the general framework of the proposed system.

- STEP1: Acquire the dataset suitable for the prediction purpose and having appropriate features.
- STEP2: Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.
  - Data cleaning
  - Fill in the missing values
- STEP3: Drawing out some important features based on the feature elimination technique and feature importance.
- STEP4: Apply machine learning algorithms (Decision Trees, Random Forest, Regression) and calculate the prediction score.
- STEP5: Implement models with a scoring tools such as log loss, grid search etc. to carry out the prediction score and feasibility of the following model.
- STEP6: Carry out the most appropriate tool and apply the model for business decisional model or prediction model.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

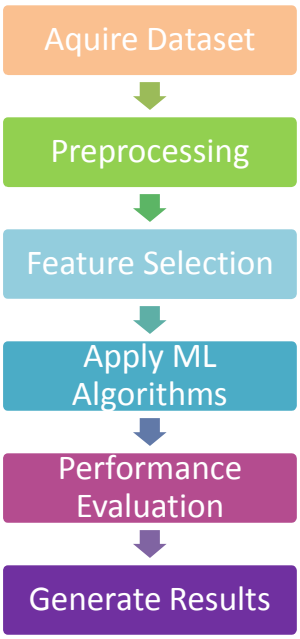

Figure 1: General Framework

4.2 Datasets

The dataset acquired for the project is a prebuilt dataset sourced from different social media sites and websites. It consists of 5000 collections of movies from all over the world (mostly UK, USA). It has following features:

Numerical Columns (16):

'num_critic_for_reviews' , 'duration' , 'director_facebook_likes' , 'actor_3_facebook_likes' , 'actor_1_facebook_likes' , 'gross','num_voted_users', 'cast_total_facebook_likes','budget', 'title_year', 'actor_2_facebook_likes' ,'facenumber_in_poster' ,'num_user_for_reviews', 'imdb_score' , 'aspect_ratio' , 'movie_facebook_likes'.

Categorical Columns (12):

'color' , 'director_name' , 'actor_2_name' , 'genres' , 'actor_1_name' ,'movie_title' , 'actor_3_name' , 'plot_keywords' , 'movie_imdb_link' , , 'language' , 'country' , 'content_rating'.

Dataset Size:

Rows - 5036
Column – 28

Target Variables:

1. Imdb_score: Numerical features ranging between 0-10 determing the imdb review scores.
2. Gross: A numeric variable having range in 10^8 and describing the box-office collection of movies.

4.3 Experimental Settings

A. Random Forest Classifier

Random forest classifier is an ensemble ml algorithm used for classification and other task that operate by constructing a multitude of decision trees at training time and outputting the class. Table I decides the Success of movie based on 3 labels i.e. (Hit, Avg, Flop).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

Target Variable– **'success'** (class of imdb_score).

	Labels	Bins
Success	Hit	1 – 3
	Avg	3 – 6
	Flop	6 – 10

Table 1: 'Success'

B. XGBoost Regression

Xgboost is the decision tree-based ensemble model that uses gradient boosting framework for user defined prediction problems.

Target variable:**'imdb_score'**
Range: 1 – 10
Type – Numeric (float)

C. Decision Tree Classifier

Decision Tree is an ML model that splits the decisional tree nodes on all available variables and then selects the split which results in most homogeneous sub-nodes. Table II predicts the gross_class of movie based on 4 labels i.e. ('C1', 'C2', 'C3', 'C4').

Target variable:**'gross_class'** (classes of gross).

	Labels	Bins
gross_class	'C1'	$100 - $100M
	'C2'	$100M - $250M
	'C3'	$250M - $500M
	'C4'	$500M - $1000M

Table 2: 'gross_class'

V. Results And Discussions

As we have used multiple ml models and algorithms for resulting different output parameters we have came up with a lot of defining results. In order to increase the accuracy, the algorithms are made to run multiple times with changed settings and factors. Let us discuss some of them here in this section.

RFECV (Recursive Feature Elimination with Cross Validation)

This technique is used to determine the importance of features with respect to the model's accuracy. Fig II shows the RFECV feature importance.

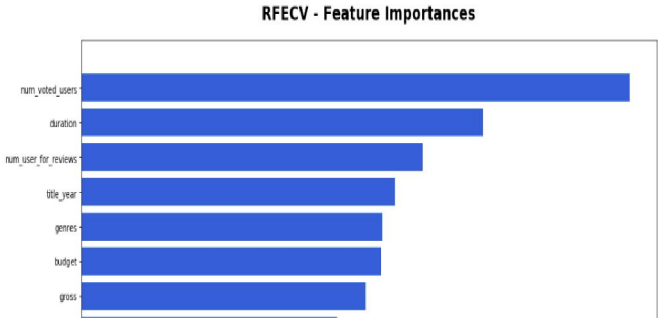

Figure 2: Feature Importance

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

5.1 Random Forest Classifier

This algorithm was used to determine whether the upcoming movie is going to be HIT, FLOP or AVG. After applying the model, the accuracy came out to be quite good which is around(80%). Below Table III shows the variation of accuracy with varying features and class density.

S.No	Test_size	HIT	AVG	FLOP	Accuracy
1.	25%	3314	1516	45	80%
2.	25%	1482	1832	1561	64%
3.	25%	3314	--	1561	78%
4.	30%	3314	1516	45	80%
5.	30%	1482	1832	1561	64%

Table 3 : Random Forest Results

5.2 XGBoost Regression

XGBoost regression model here results out the review of a movie by learning from historic movie dataset. This regression model is targeted on a feature with numeric values. Using it with grid_search algorithm makes it even more accurate. Grid_search algorithm takes significant amount of time for implementation and hence makes the prediction precess a little time-taking. Statistical parameters and accuracy of the model with respect to different features number and test size is calculated and shown in the Table IV:

No of features (Test size)	Grid_search	Mean Absolute Error	Model Score	R2 Score	Accuracy
6 (25)	No	0.57	0.914	0.483	89.16%
	Yes	0.55	0.752	0.508	89.34%
6 (30)	No	0.58	0.924	0.471	89.18%
	Yes	0.56	0.671	0.509	89.39%
All (25)	No	0.55	0.974	0.525	89.57%
	Yes	0.52	0.842	0.563	89.9%
All (30)	No	0.56	0.977	0.515	89.57%
	Yes	0.52	0.847	0.56	90.04%

Table 4: XGBoost Results

5.3 Decision Tree Classifier

To predict the box-office collection, the gross feature is distributed in classes and decision tree classifier is used to predict the collection class(gross_class). Principal Component Analysis (PCA) was also used for better accurate results but inspite of that the accuracy considerably fell down. It has 2 criterion – 1. gini 2. entropy. Table V shows the best possible classification report the model has obtained:

	Precision	Recall	f1-score	support
C1	0.49	0.72	0.59	262
C2	0.00	0.00	0.00	145
C3	0.75	0.78	0.77	577
accuracy			0.68	984
Macro avg	0.41	0.56	0.55	984
Weighted avg	057	0.68	0.66	984

Table 5: Classification Report

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

Table VI is comparison of Decision Tree Accuracy Vs Classes with and without Grid Search.

Classes	Accuracy	
	Without Grid Search	With Grid Search
3	65%	68%
4	51%	57%
5	43%	48%

Table 6: Decision Tree Accuracy Vs Classes

VI. CONCLUSION

Thus, we could comprehend that the prediction of movie success is certainly possible with high percentages of accuracy. So by using our prediction engine of our application, production house can evaluate before it production begins. And production house can decide that movie is worth investing or not and accordingly make their decisions. We can expect that with valid accurate data and fuzzy logic is an effective means of categorizing predictions and adds more accuracy and dynamicity to the system. In case of multiple cast like more actor, actress, director and writer this current application is not suitable· enough. Here we can just one actor, actress and director as input.

The predictive model will predict the movie success based on historical movie events and their success rates. For future work, we can add trailer hit count and social media hive, news trending and release date, vacation of years to this system to get the system more mature for prediction.

REFERENCES

[1] Steven Yoo, Robert Kanter, David Cummings TA, Andrew Mass, "Predicting Movie Revenue from IMDB Data", pp. 1-0 ,2015

[2] Nevon Project "Movie Success Prediction Using Data Mining" [Online] Available: http://nevonprojects.com/movie-success-prediction-using-data-mining/ [Accessed: 10Dec-17]

[3] T. G. Rhee and F. Zulkernine, "Predicting Movie Box Office Profitability: A Neural Network Approach," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016

[4] Karl Persson ECTS, "Predicting movie ratings, A comparative study on random forests and support vector machines", pp.1-28, 2015

[5] Saurabh Kumar , " Movie Success Prediction using Data Mining For Data Mining and Business Intelligence (ITA5007) of Master of Computer Application School Of Information Technology and Engineering " , April 2019.

[6] Nahid Quader , Md. Osman Gani, and Dipankar Chaki, "Performance Evaluation of Seven Machine Learning Classification Techniques for Movie Box Office Success Prediction" , in 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), 7-9 December 2017, Khulna, Bangladesh

[7] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in textregression. In Proceedings of NAACL-HLT, 2010.

[8] N. Darapaneni et al., "Movie Success Prediction Using ML," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020

[9] Vr, Nithin & Pb, Sarath. (2014). Predicting Movie Success Based on IMDB Data.

[10] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,"IEEE Trans. Knowl. Data Eng., pp. 734–749, 2005.

[11] [A. Japhne and R. Murugeswari, "Opinion Mining based complex polarity shift pattern handling for improved sentiment classification," 2020 International Conference on Inventive Computation Technologies (ICICT), 2020