Movie Box-office Success Forecaster

Bhavishi Bansal	Pragya Singh	Praveen Singh Samota	Vardhana Sharma
SP22001	2017305	2020104	SP22003
bhavishi22001@iiitd.ac.in	pragya17305@iiitd.ac.in	praveen20104@iiitd.ac.in	vardhana22003@iiitd.ac.in

1. Abstract

The silver screen industry is a billion-dollar industry with the extensive involvement of various stakeholders. In 2019, the global box office was worth $42.2 Billion. The movie industry is a massive sector for investment but larger business sectors have more complexity, and it is hard to choose how to invest. Furthermore, significant investments come with more significant risks. Through this project, we want to explore ML models that help us successfully predict the box office performance of movies.

2. Introduction

Movie revenue depends upon many factors, some of which are the cast, the budget and the reviews of the audience. A good prediction can help the production companies to have some idea of the return on investment, helping the theatre chains as well. The definition of a 'successful' movie can be relative, for someone it can be the gross profit it earned, and for others, it can be the good critic reviews and popularity among the audiences. In our research paper, we will consider the success of a movie based on its gross income only. While it's difficult to selectively consider one of these factors as the deciding factor, we can develop a Machine Learning model to solve our problem.

3. Literature Survey

1. The paper [1] proposes a decision support system for the movie investment sector using machine learning techniques such as Support Vector Machine (SVM), Neural Networks and Natural Language Processing; the system predicts a movie box office profit based on some pre-released features and post-released features. The target class is divided into five categories based on the revenue the movie generates. The 5th target class has the highest profit (>150M) and the least target class value is 1, for those movies which generate less than 0.5M profit. The authors examined a lot of online factors that influence success, like IMDb rating, Tomato rating, review sentimental values, etc. Sentimental analysis is done using a text analytics API. The authors concluded that besides the pre-release and post-release factors, some inconsiderable details like the GDP of a country at the time of the release, and the number of tickets sold in a particular year, can affect the success of a movie as well. The prediction of sequels using these features and modeling is inefficient, as a sequel may or may not work based on the performance of its prequel.
2. This paper [2], builds a prediction model using machine learning techniques such as random forest classifiers, decision trees, and regression models. The authors examined a lot of previous work on this problem and arrived at a point to include new features to increase efficiency. Social media likes are one of the highlights of this paper; facebook numbers of directors, actors and movie pages are considered as features. Instead of focusing on the gross income of the movie as a target value only, the authors considered the IMDb score as well. Various models like XGBoost Regression, Random Forest Classifier, and Decision Tree Classifier are used and their results are discussed. The authors concluded that a good prediction model is possible, with more accurate data. In the case of multiple actors, and directors, this model will fail as it considers only one actor/director. For future work, the authors suggested using post-release features like trailer counts as well to get better efficiency.

4. Dataset: Dataset details with data preprocessing techniques

1. Data acquisition: The dataset contains 5043 movies retrieved from the IMDb dataset present on Kaggle. The features are:- color, movie title, director name, critic numbers for reviews, duration, director's facebook likes, actor 1/2/3 facebook likes and names, gross revenue generated,number of voted users, facenumber in poster, plot keywords, movie imdb link, number of user reviews, language, country, content rating budget, year of release, aspect ratio, movie's facebook page likes, and IMDb score.

2. Data cleaning:- Duplicate entries were deleted, and numerical and categorical columns were identified to treat them separately. In categorical features, null values were recognized, first of all, several color values were null, and they were replaced with the mode value of the feature. Rows with empty actor names were identified with their titles and replaced with the appropriate names we found using the internet. In the same way, we filled up the values of empty language and country rows. After that, the rows still left with null values in categorical columns were dropped. Similarly, the null values were checked for numerical features. Duration got only one row with empty value, we replaced it with the duration time available on the IMDb page of the movie.

3. Data processing:- Used label encoder from Sklearn library to label encode the categorical columns, We created a correlation plot to depict what entities are most related to the revenue generated and the IMDb score. Leading it into data visualization, we plotted the graph of Facebook Likes vs Year the movie got released to determine the popularity of movies in general over the years.We created a doughnut pie chart for the content rating, a bar graph for the IMDb score count (Count vs. IMDb score), a bar graph to denote the top 10 movies based on cast popularity (movies vs Facebook likes), a pie chart for distribution of movies among decades, lineplot for IMDb score vs duration, and a few bar graphs for top movies based on various factors like budget, gross, movie Facebook likes, etc.

5. Methodology

Our Goal is to build a machine-learning model to predict the success/revenue of a new movie given such features as budget, release dates, and genres. A conclusive note based on model training and testing process with some detailed insights on the tested data.
We decided to create models based on two different features, i.e. IMDb score and gross revenue generated. The reason is, a movie can be successful in one of two terms, it doesn't need to be financially successful to get a good audience rating. IMDb scores reflect how much the critics and audience accepted the movie. At the same time, the revenue depicts how well the movie did 'financially' among the general audience, irrespective of how the critics took it. Both parameters are equally important.
For the IMDb score, we first bucketized the dataset into three categories: low, medium, and high.

6. Results and analysis

Implemented various models - logistic regression, random forest, gradient boosting, ada boosting, decision tree (gini and entropy), Naive Bayes, Support Vector Machine, and KNeighbors Classification.

The results for Gross were better than those for the IMDb score. The highest accuracy achieved for IMDb was 0.7914 using Random Forest, while for models using gross 0.9451 using Random Forest here as well. After training the random forest model, we analyzed the importance of each feature in predicting the target value and concluded that the budget has the highest importance in deciding the results.

7. Conclusion

Accuracy was an issue for the IMDb score models, which depict an entirely different picture than the assumption we made during the implementation phase. There's a vast difference between the average accuracy of models using gross as the target variable and models using IMDb score as the target variable.

8. References

[1] N. Quader, M. O. Gani, D. Chaki and M. H. Ali, "A machine learning approach to predict movie box-office success," 2017 20th International Conference of Computer and Information Technology (ICCIT), 2017

[2] Jatale, Sanyam & Moze, Rohan & Khandekar, Varsha & Jain, Shubham & Mokate, Sanket. (2021). Analyzing and Predicting The Success of Box Office Collection of a Movie Using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology.

[3] Medium. 2022. How to use Machine Learning Approach to Predict Movie Box-Office Revenue / Success? https://medium.com/analytics-vidhya/how-to-use-machine-learning-approach-to-predict-movie-box-office-revenue-success-