**MOVIE SELECTION AND STUDIO AQUISITON USING IMDB DATASET**

Felix Kimeu, Ruth A. Frimpong, Charles Owusu-Ansah and Brandon A Obasogie

**College of Science, Engineering and Technology, Nor-folk State University, Norfolk, 23504, Virginia, USA**

[f.m.kimeu@spartans.nsu.edu; r.a.frimpong@spartans.nsu.edu; [c.owusu-ansah@spartans.nsu.edu](mailto:c.owusu-ansah@spartans.nsu.edu); b.a.obasogie@spartans.nsu.edu]

**ABSTRACT**

For companies to acquire subsidiaries they base their decision on data. This project intends to help a movie industry to select profitable studios to purchase and also to enable users to pick movies to watch based on filtering system that makes it easier to come up with movie choice. Most works focus on either the business or entertainment aspects on movies. However, this work focused on using data set analysis to produce result suitable for those who seek to do business and those who seek entertainment. Algorithms that calculate profits and suggest movies based on user inputs were proposed. These algorithms provide information necessary for major studios to acquire subsidiaries and also enable a user to pick movies to watch based.

**Keywords:** IMDb dataset; Data Analytics; Analytics; Dataset; Dataframe, user ratings.

**INTRODUCTION**

Getting the right movies to watch is sometimes tedious. While some end up giving up watching movies out of frustration after wasting several minutes deciding, others do not even attempt to login onto sites like Netflix if they don’t know before-hand the kind of movies to watch. A filtering system that takes two inputs from users and then give them movies suggestion is highly recommended.

To form a subsidiary, companies must hold a board meeting and vote on the decision by indicating the type of business entity that needs to be selected. Management cannot finalize decisions without strong basis of the company’s profitability.

Studies have been done to on how news data and IMDB data can be used in predicting movie gross, as well as in predicting movie ratings [1, 2]. Machine learning has also been applied in the prediction of movie gross in past studies [3].

Our analysis focuses on using the movie gross and the production budget to calculate profits to give companies insights on the subsidiaries to acquire. We also focused on creating an algorithm to suggest movies based on a meta-score.

**LITERATURE REVIEWS**

Yoo et al. created regression and classification models for comparison purposes. Their study found that linear regression and logistic regression worked in a similar manner in data classification. However, linear regression had a better correlation with the actual gross revenues [1]. The study focused on a number of features including simple numeric, text, and sentiment features [1].

Zhang et al. used quantitative news data and created a system for large-scale news analysis to improve movie gross prediction [2]. Two models – regression analysis and k-nearest neighbor model were tested. In this study, models using news data were able to achieve results similar to those using IMDB data. A combination of IMDB data and news data resulted in a statistically significant improvement in the prediction performance [2].

A study by Ahmad et al. focused on the different sources of data, the techniques, the features, and the evaluation metrics used in movie revenue prediction [3].

**METHODOLOGY**

Fig1 show the flow of our approach.

Figure . Process flowchart

Diagram

Description automatically generated

IMDb dataset was acquired from Kaggle

The acquired dataset was visualized to understand and select the correct methods for our analysis. The data was then cleaned by dropping NAN, removing currency signs and performed other data pre-processing methods. To identify the studios and the movies producing the highest profits, the profit for the individual studios, and the movies, was calculated as shown in the equation below:

profit=gross-budget

This calculation only took into account the production budget.

The top 10 profitable movies and the top 10 profitable studios were plotted, as well as the 10 studios which made the greatest losses.

A screening criteria was developed to help users select movies to watch based on user input. The aim of the screening criteria was to suggest movies for the users based on whether the users provide both year and genre, either year or genre, or whether the user does not provide any input.

A meta-score was developed as a means to rate the movies and bar charts of the highly rated movies, and the least rated movies were plotted.

Meta-score = the mean of normalized (Votes, user reviews and critics review)

**FINDINGS AND ANALYSIS**

Figure 2 below shows the top 10 movies in terms of profit. The profit column for the movies and the studios was created by calculating the profit as shown in the equation below:

An important take away from the plots in this section is that having one profitable movie does not necessarily makes a studio most profitable. Consistent with data from different sources in 2019, we identified Avatar and Avengers: Endgame to be making the highest profits.

**Chart, bar chart

Description automatically generated**

Figure . Top movies

The data was also grouped by the studios and each studio’s profit/loss was calculated. The data was plotted as shown in figure 3 and figure 4, where figure 3 represents the top 20 studios in profits and figure 4 represents the studios making the highest losses.

Chart, bar chart

Description automatically generated

Figure . Top 20 studios

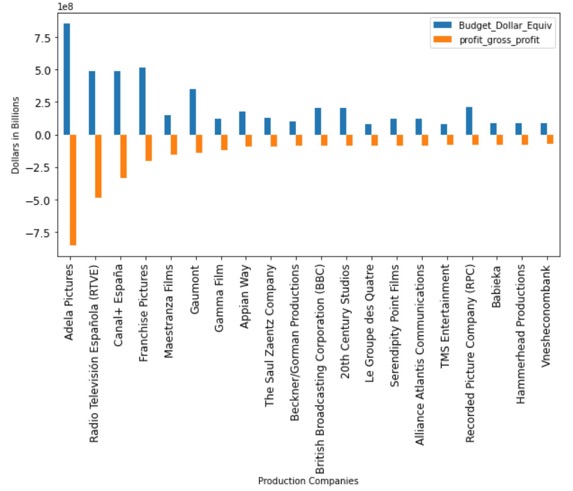


Figure . Bottom 20 studios

To suggest movies for users, we created a metascore which ranked movies based on user input. The metascore was based on normalized data on votes, user reviews and critics review.

Figure 5 below shows the combination of inputs which the users could provide in the movie suggestion algorithm, and an output from the algorithm.



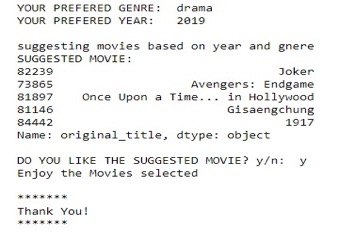


Figure . User inputs and Algorithm output

**CONCLUSIONS**

The proposed algorithms discussed will give companies leadership adequate data to decide which studios to acquire. In addition, the data sets used for the algorithm can be enhanced/ refined with up to data metrics to give a better gauge of profits per fiscal quarter. With updated data the algorithm would essentially readjust and display the top profitable studios with updated budgets and profit data.

The filtering system our group introduced to enable users to pick movies based on their preference was effectively implemented. This system will give users the ease to choose movies by more than just genre and type of media ( TV shows or movies).

Lastly, the created metascore to gauge a movies appeal and success. Using users votes and reviews combined with critic reviews the created metascore became the projects standard for a successful movie. The metascore concept is already used in industries but future work can be done to make its efficacy more validated.

In closing, our group did successfully clean a dataset and manipulate the data to reach the desired goals. The result can aid companies in the acquisition of profitable and failing studios.

**REFERENCES**

[1] S. Yoo, R. Kanter, D. Cummings, and T. A. Maas, Predicting Movie Revenue from IMDb Data. .

[2] W. Zhang and S. Skiena, “Improving Movie Gross Prediction through News Analysis,” in 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Sep. 2009, vol. 1, pp. 301–304, doi: 10.1109/WI-IAT.2009.53.

[3] I. Said Ahmad, A. Bakar, M. R. Yaakub, and S. Hassan Muhammad, “A Survey on Machine Learning Techniques in Movie Revenue Prediction,” *SN Comput. Sci.*, vol. 1, Jul. 2020, doi: 10.1007/s42979-020-00249-1.