Proceedings of the Fourth International Conference on Electronics and Sustainable Communication Systems (ICESC-2023)
IEEE Xplore Part Number: CFP23V66-ART; ISBN: 979-8-3503-0009-3

2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) | 979-8-3503-0009-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICESC57686.2023.10192928

A Method for Predicting Movie Box-Office using Machine Learning

Menaga D
Department of Computer Science and Engineering
St. Joseph's Institute of Technology
Tamil Nadu, India
sjit.csedept@gmail.com

Akshaya Lakshminarayanan
Department of Computer Science and Engineering
St. Joseph's Institute of Technology
Tamil Nadu, India
akshayal365@gmail.com

***Abstract*— Movies are still an important source of entertainment in any country. But the industry loses a lot when the movies don't do well at the box office. A new area of data analysis is predicting how society will respond to new items in terms of popularity and adoption rates. This project helps analyze and predict the theatrical performance of films before they are released independently. When a movie becomes a blockbuster, it makes a huge profit. But when the movie fails, the loss is also huge. And both have direct positive and negative effects. This also helps investors associated with this business avoid investment risks. Researchers show that it accounts for almost 25% of the movie's revenue. Within 1-2 weeks after release as such, it is difficult to predict how a film will perform at the box office in advance of its release date. The proposed system attempts to predict the success rate of movies. Predictive analysis of various properties of the film. The film attributes such as crew, release date, Production Company, and plot are valuable sources of information and can be of great help in predicting the success of a film. There is a lot of data available on the internet from various sources such as TMDB about different movie attributes, and it is highly relevant and successful, making it an important use case for machine learning.**

Keywords*— *Movies, Box-office, Machine Learning, Revenue Prediction, Regression.

I. INTRODUCTION

As technology rapidly advances the production of films is sped up by the use of digital cameras and other technology in comparison to classic film cameras. More people than ever before are going to the cinema. As a result, it can be said that the cinema industry is one of the most important segments of contemporary society. Every movie has the potential to represent society and influences viewers' opinions. A good movie may inspire, inform, and delight its viewers in a variety of ways. A film has the power to arouse our interest, teach us about a different culture, change our perspective, or introduce us to a previously unknown aspect of the world. The effect that movies have on people should beconsidered. Every movie has a distinct culture in which it is set and produced. Inseparable from our daily lives, movies have become.

Box office revenues have steadily become an increasingly major source of income over the last few decades, during which time the motion picture industry has been seeing sustained and robust growth. It has never been easy for production firms to make precise projections about the financial success of their films until they have been released in theaters. Despite this, studios are always on the lookout for fresh methods that will enable them to improve their predictive powers.

As a result of recent developments in machine learning and big data analytics, it is now feasible to design models that precisely predict the amount of money that a movie will produce at the box office. This is an exciting development for the entertainment industry. The entertainment industry is going to be quite excited about this new turn of events. In this investigation, the use of machine learning to solve the challenge of accurately predicting the level of financial success that will be achieved by blockbuster movies at the box office.

Following the amassing of these data points, a machine learning model is then allowed to be trained to make forecasts regarding the monetary performance of movies at the box office. For the model to be able to provide an accurate prediction of how well a movie will perform at the box office, it must first be trained using an enormous amount of data that has been gathered in the past. It is impossible to overstate how important it is for a studio to have accurate box office projections when it comes to being able to influence the business decisions that are made by the studio.

Studios can more effectively manage their resources, fine-tune their marketing, and make more educated judgments on future projects when they can precisely predict the amount of money a film will produce at the box office. This allows the studios to make more informed decisions regarding which projects they will pursue in the future. By using machine learning and analytics to massive volumes of big data, studios may be able to gain a competitive advantage and increase their capacity to forecast the commercial success of their films.

II. RELATED WORK

Movies have a significant impact on the way our world works. However, nobody can accurately forecast how well a movie will do at the box office. The motion picture industry is now seeing annual growth of 11.5%. [1] While some films with a large budget are unsuccessful, others with a lower budget are great box office draws. [2] The

cast, crew, posters, storyline elements, budget, production companies, release dates, languages, and countries are all taken into consideration when estimating how much money a film will make at the worldwide box office as a whole. [3] Every movie has the potential to both reflect and change the views of the society it's a part of. When a movie is a "Blockbuster Hit," not only are the earnings enormous, but they are also important. [4] And each factor directly influences the consequences, whether they are favorable or bad. [5] A well-made film may provide a variety of benefits to its viewers, including amusement, education, and motivation. [6] The development of learning algorithms that can be used by machines and do not need participation from humans is the goal of this project. [7] Machine learning methods have created better prediction results than the previous strategies owing to machine learning's efficient utilization of data. [8]

As artificial intelligence develops; more and more individuals are getting interested in utilizing machine learning techniques to estimate movie ticket sales. [9] This is because machine learning has produced better prediction outcomes than the earlier techniques. [10] There is a strong correlation between the daily audience counts, the prospective revenues of the movie at the box office, and the audience approval ratings. With the help of a variety of data analysis methods, one may estimate movie ticket sales. [11] Because it contains all of the crucial information that is necessary for prediction, the Kaggle dataset may be utilized to forecast how lucrative a film will be at its theaters. [12] It is essential to do data analysis to interpret and investigate the data in relevant ways. Data analysis is the process of categorizing, interpreting, organizing, and presenting data to make sense of the data and offer context for the data. Regression analysis is one of the strategies that may be used to anticipate the corresponding response. [13] The statistical technique known as regression connects one or more variables that are independent of a variable that is under study. The accuracy of the models was evaluated based on the testing data, while the prediction model was developed with the help of the training data. [14] The goal of this study was to determine which of several different algorithms produces the lowest mean absolute error. The mean absolute error (MAE) is a measure of how far off the observed value is from the value that was predicted for it. [15] The model is the product of an in-depth investigation of the ML algorithms that locate successful movies using the TMDB dataset. These algorithms include Random Forest, Decision Tree, and XGBoost.

III. PROPOSED WORK

The objective of this research is to identify a machine learning model that, when fed with the information that has been provided, is capable of making accurate projections about the amount of money that a movie will generate at the box office. It is anticipated that the algorithm will have the ability to provide an accurate prediction of the level of success that the movie will have when it is made available in cinemas. During the training process, additional data ought to be added so that the model can acquire accurate information. In addition to this, the precision of the myriad of methods that have been used in the classification of the data into categories, investigation analytical strategies that have the potential to enhance precision, and evaluate the precision of these many ways. The output of putting this model through its paces may be put to use in the form of an educated guess on a film. The following structure demonstrates how the method for the prediction model should be carried out: shown in Figure 1.

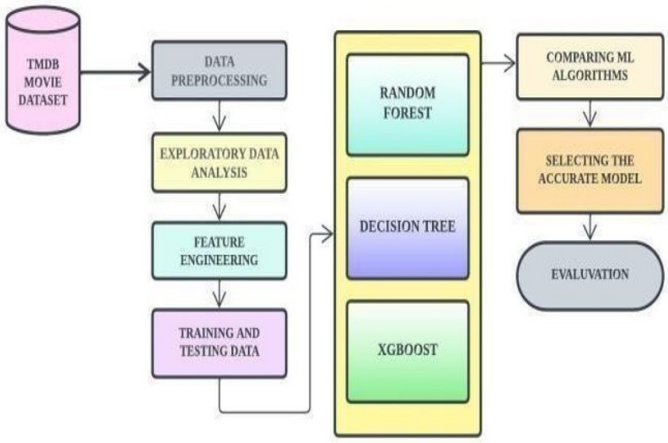

Fig 1. Proposed Model

The modules are

Module 1: Data Pre-processing
Module 2: Exploratory Data Analysis
Module 3: Feature Engineering
Module 4: Training and Testing Data
Module 5: Evaluation

A. Data Preprocessing

Because the level of data quality and the amount of information that can be gleaned from it has a direct bearing on the ability of this model to learn, the process of data preparation is an essential step in the overall machine learning process as shown in figure 2. To get the data ready for analysis, it discusses how to clean the data, transform it, and integrate it. Because of this, it is very necessary to do any necessary preprocessing on the data before providing it to our model. It's conceivable that the dataset could on occasion be missing some data, including some category data, or something else entirely. To enhance the quality of the data, this will be subjected to cleaning and modification.

Fig. 2. Missing values in the data

B. Exploratory Data Analysis

A step during the evaluation of the data process is called exploratory data analysis (EDA). Making sense of the data may have at this point can help may determine what questions to ask, how to frame them, and the best way to change the data may have to acquire the answers may need. by combining visual and quantitative tools to gain an understanding of the story this conveys while taking a comprehensive look at patterns, trends, outliers, unexpected outcomes, and other factors in existing data. With a few exceptions, most EDA approaches in Figure 3 are graphical to understand the dataset better.

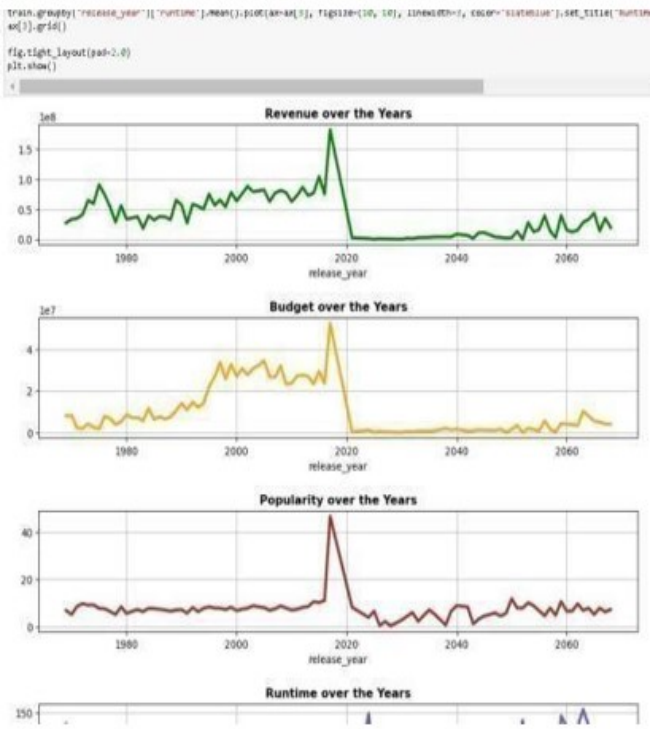

Fig. 3. Annual Analysis EDA

C. Feature Engineering

The machine learning process begins with the stage known as feature engineering, which enables the extraction of features from raw data as shown in Figure 4. These features may then be used in the succeeding step of developing predictive models. After gathering this information, one may utilize it to create predictions. In a general sense, all algorithms that are employed in machine learning make use of input data to create output. This is because input data is necessary for the algorithms to function properly. The input data is still displayed in the form of a table, with rows standing for specific occurrences or observations and columns representing variables or characteristics; these variables or characteristics are also sometimes referred to as features. The word "feature" relates to the topic that is being addressed in this article, and the definition of a "feature" in this context is a characteristic that affects or helps with an issue.

Utilizing various feature engineering strategies is one way in which the accuracy of the model can be brought up to a higher level. It takes the data that is supplied and converts it into a format that may be comprehended in a shorter amount of time. Within the context of this particular circumstance, one of our objectives is to enhance the degree of openness that the model of machine learning possesses.

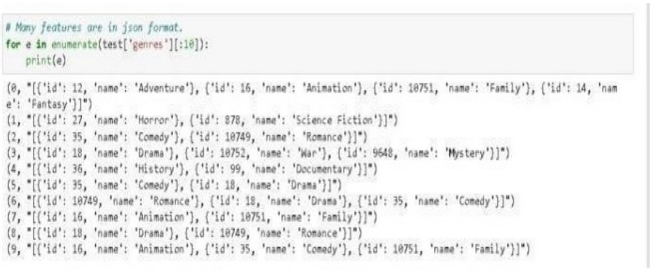

Fig. 4. Enumerating the JSON Format Features

D. Training and Testing Data

Machine learning predictions heavily depend on the grade of the data, and if not feeding this model with the proper data, it won't provide the desired outcome. In projects involving machine learning, the original data is typically separated into evaluated data and data for training. The proposed work test whether the model generalizes effectively to the new or unknown dataset or test set after training it on a subset of the original dataset, or the training dataset. Once the model has been trained using train it on a portion of the training dataset, which is the initial data set. After being trained on the training dataset, the model should now be put to the test on the test dataset. This dataset assesses the model's performance and validates that it generalizes well to fresh or unproven data.

Fig. 5 Testing and training data

E. Evaluation

Figure 5 shows the training and testing data. After training is finished, evaluation should be used to determine the model's effectiveness. This is the point at which the dataset previously set aside is useful. In the model, different ML methods are utilized to compare the model's accuracy.

- Random Forest: The class with the most votes will ultimately be considered as the predictor's output in the group learning technique known as Random Forest, which is employed in classification and regression.

- Decision Tree: The decision Tree algorithm is used in the model which divides a node into two or more sub-nodes while increasing the homogeneity of the resulting sub-nodes. It is a classifier with a branching structure, where each leaf node provides the classification result and the interior nodes contain the data's features.

- XGBoost: With XGBoost, a distributed gradient boosting library that has been enhanced, the residual trees are generated by comparing the scores between

the leaves and the nodes that came before them to choose which variables to utilize as the roots and nodes.

Because it is a regression model, it is only able to provide accurate forecasts for values that are either higher than or lower than the actual figure. As a direct result of this, the only choice available is to get residuals. It is possible to provide a straightforward definition for the MAE by stating that it is equal to the total residual divided by the total number of points in the dataset. The aforementioned algorithms are compared with one another to ascertain whether one of the algorithms generates the model that can anticipate movie profits with the greatest degree of precision.

IV. RESULTS

When compared to other machine learning algorithms, the random forest approach yields results that are accurate and reliable, and the mean absolute error is used to evaluate the approach's level of precision. Based on the information that was included in the dataset relating to the movie, the algorithm was able to provide an accurate forecast of the performance of the film. Figure 6 shows the mean absolute error of the algorithm and Figure 7 depicts the revenue of the movies.

```
# Prediction
y_pred_rf = rf_model.predict(X_valid_full)

# Calculate MAE
mae_rf = mean_absolute_error(y_pred_rf, y_valid)

print("Mean Absolute Error RF:" , mae_rf)
```

```
Mean Absolute Error RF: 1.4121268916093035
```

Fig. 6. Mean Absolute Error of the algorithm

```
: output.head()
```

	id	revenue
0	3001	12.402519
1	3002	13.345052
2	3003	15.830939
3	3004	14.345368
4	3005	12.994582

```
: print(output['revenue'])
```

```
0       12.402519
1       13.345052
2       15.830939
3       14.345368
4       12.994582
          ...
4393    18.003410
4394    17.042542
4395    17.178644
4396    16.145921
4397    14.479192
Name: revenue, Length: 4398, dtype: float64
```

Fig. 7. Revenue of the movies

Advantages:

The predictions that are produced by a random forest are reliable and easy to comprehend on their own. The management of very massive datasets is not an impossible obstacle to overcome. When compared to the decision tree approach, the random forest algorithm provides a higher degree of accuracy in the predictions it makes on the outcomes of events. This is because the algorithm takes into account more variables. Both approaches are utilized to make projections regarding the outcomes of scenarios. Ideal if the following particular circumstances into account:

1) Estimate the box office take.
2) Powerful forecasting method
3) Helpful for precise forecasting of which films need to bereleased first.
4) Reliable and secure system

Disadvantages:

By taking a look at the country's rate of gross domestic product (GDP), one can determine whether or not there is a level of financial stability in a nation at the time that a movie is being produced. When times are tough economically, a minuscule percentage of the general public will still make the effort to go to the movies to satisfy their need to see motion pictures. As a result, the criteria discussed above have a significant role in determining a movie's overall level of success. As a direct consequence of this, it is impossible to forecast what will take place.

Limitation:

The number of people who see a movie is essential to its overall success. If there is no one viewing movies, then the whole business is pointless since that is what it is all about. If there is no one watching movies, then the entire industry is meaningless. The number of tickets that were purchased during a certain year may provide some information about the size of the audience during that year. Nevertheless, the role of moviegoers is influenced by a variety of circumstances, such as the political climate and economic stability of a nation. It is possible to tell whether or not there is financial stability in a country at the time that a film is released by looking at the GDP rate of that nation. It is possible to tell whether or not there is financial stability in a country at the time that a film is released by looking at the GDP rate of that nation. When things are tough economically, a much less percentage of individuals will go out to the movies. These aspects, therefore, are essential to the final success of a movie. Therefore, in this scenario, our model is unable to forecast the amount of revenue.

V. CONCLUSION

In most instances, the methodologies section is then followed by a summary of this work that is arranged logically. The information in this article is provided in a logical sequence, and in contrast to other algorithms and optimization methods, it has shown effective outcomes for prediction. The purpose of this study is to assess the effectiveness of the aforementioned models and to do so, the proposed work will be making use of machine learning techniques. The research also demonstrates that the random

forest model performs better than other methods of machine learning because it generated a lower mean absolute error (1.41) than the other methods, which is another sign of performance. This is demonstrated by the fact that the mean absolute error generated by the random forest model was 1.41, which was lower than the mean absolute error generated by the other methods. The application of the Random Forest algorithm to box office records enables one to make an accurate forecast of the amount of money a movie will make at the box office.

REFERENCES

[1] T. Sharma, R. Dichwalkar, S. Milkhe, and K. Gawande, "Movie Buzz - Movie Success Prediction System Using Machine Learning Model." 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 111-118.
[2] N. Darapaneni, "Movie Success PredictionUsing ML," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2020, pp. 0869-0874.
[3] G. Velingkar, R. Varadaraian, S. Lanka, and A. K. M, "Movie Box-Office Success Prediction Using Machine Learning," 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India,2022, pp. 1.
[4] H. Li, "Using machine learning forecasts movie revenue," 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Hangzhou, China, 2021, pp. 455-460.
[5] Jiang and Z. Wang, "Predicting box office and audience rating of Chinese films using machine learning," in Proceedings of the 2018 International Conference on Education Technology Management, 2018, pp. 58–62.
[6] Timani, P. Shah and M. Joshi, "Predicting Success of a Movie from Youtube Trailer Comments using Sentiment Analysis," 2019 6th International Conference on Computer Science to develop global sustainability (INDIACom), New Delhi, India, 2019, pp. 584-586.
[7] S. J. J. Rathnayaka, C. J. Ranathunga, R. Navarathna, A. Kaneswaran, and Y. Balathasan, "Predicting Movie Ratings from Audience Behaviors on Movie Trailers," 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS), Negombo, Sri Lanka,2021, pp. 489-493.
[8] H. Timani, P. Shah and M. Joshi, "Predicting Success of a Movie from Youtube Trailer Comments using Sentiment Analysis," 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2019, pp. 584-586.
[9] W. Lu, "Research on Prediction of Movie Box Office Based on Internet Comments," 2019 12th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2019, pp. 11-14.
[10] A. Pocol and L. Istead, "Assessing the Impact of Movie Plot Summaries on Box Office Sales," 2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA, 2022, pp.48-52.
[11] 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.
[12] Y. J. Kim, Y. G. Cheong, and J. H. Lee, "Prediction of a movie's success from plot summaries using deep learning models", *Proceedings of the Second Workshop on Storytelling*, pp. 127-135, August 2019.
[13] R. Yao and J. Chen, "Predicting movie sales revenue using online reviews", *2013 IEEE International Conference on Granular Computing (GrC)*, pp. 396-401, 2013.
[14] Y. Liu, X. Huang, A. An and X. Yu, "ARSA: A sentiment-aware model for predicting sales performance using blogs", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval ser. SIGIR '07*, pp. 607-614, 2007.

[15] H. Sadadi, D. Aloufi, and Z. Ye, "Predict movie revenue by sentimental analysis of Twitter", *Proceedings of the International Conference on Data Processing and Applications ser. ICDPA 2018*, pp. 1-4, 2018.