# BOX OFFICE PREDICTION

*A*

*Mini Project Report Submitted in partial fulfilment of the*

*Requirements for the award of the Degree of* **BACHELOR OF ENGINEERING** IN

### INFORMATION TECHNOLOGY

By

SYED MUNEEBULLAH-1602-19-737-178 CH. SREEYA - 1602-19-737-176 S.SHRAVYA-1602-19-737-170



### Department of Information Technology Vasavi College of Engineering (Autonomous) (Affiliated to Osmania University) Ibrahimbagh, Hyderabad-31

**2020**

**Vasavi College of Engineering (Autonomous) (Affiliated to Osmania Universit**y) **Hyderabad-500 031**

**Department of Information Technology**



**DECLARATION BY THE CANDIDATE**

We, **Syed Muneeb, Ch. Sreeya** and **S.Shravya**, bearing hall ticket numbers, **1602-19-737- 178,1602-19-737-176** and **1602-19-737-170**, hereby declare that the project report entitled "BOX OFFICE PREDICTION” is submitted in partial fulfilment of the requirement for the award of the degree of **Bachelor of Engineering** in **Information Technology.**

This is a record of bonafide work carried out by us and the results embodied in this project report have not been submitted to any other university or institute for the award of any other degree or diploma.

SYED MUNEEBULLAH 1602-19-737-178

CH. SREEYA 1602-19-737-176

S. SHRAVYA 1602-19-737-170

# ACKNOWLEDGEMENTS

The project work in this report is an outcome of continual work and intellectual support from various sources. It is almost impossible to express adequately the debts owed to many persons who have been instrumental in imparting this work a successful status. It is however a matter of great pleasure to express my gratitude and appreciation to all those people who had helped in completion of this project.

We take this occasion to thank God, almighty for blessing us with his grace and taking our endeavor to a successful culmination. We would like to take the opportunity to thank Mr. Dharma Reddy, Vasavi College of Engineering, for giving us an opportunity to work on this project, which not only has increased our awareness, but also taught the importance of team work, it is because of his guidance, constant encouragement and inspiration that we have been able to accomplish the task of completing the project.

Our special thanks to Information Technology Department, VASAVI COLLEGE OF ENGINEERING, IBRAHIMBAGH for its earnest efforts and guidance throughout our project work. Last but not the least, we would like to thank friends for the support and encouragement they have given us during our work.

# INDEX

1. Abstract
2. Introduction
3. Proposed Work
4. Results
5. Discussions and Future work
6. References

**ABSTRACT**

Movies have become a major form of digital entertainment since the 20th century. Most movies from major production companies generate profits above $100m even after factoring in marketing costs. The film industry is a huge commercial hub now and as in any other industry, it requires planning and strategy for maximum profit.

While some movies may make money even in the billions, other movies may fail to turn up a profit. It is not always predictable as to which movies will flop and which movies will be successful. However, in the growing film industry, it is imperative to plan ahead and analyse which different factors, strategies and approaches work best in making a movie successful.

Predicting society's reaction to a new product in the sense of popularity and adaption rate has become an emerging field of data analysis. The motion picture industry is a multi-billion- dollar business, and there is a massive amount of data related to movies is available over the internet. This study proposes a decision support system for movie investment sector using machine learning techniques. This research helps investors associated with this business for avoiding investment risks. The system predicts an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic. Using Support Vector Machine (GBR), Neural Network and Natural Language Processing the system predicts a movie box office profit based on some pre-released features and post-released features.

The Box Office Prediction Project aims to solve the above problem by predicting the Box Office performance of a movie based on data such as its cast, director, crew, production company, promotion strategies, releease dates, languages, region, plot keywords etc. This Project analyses the performance of previous relevant movies to predict the revenue generated by upcoming movies. The Project aims to learn which strategies and factors are most important in making a movie successful and profitable.

# INTRODUCTION

The movie industry is one of the first forms of industrialized mass-entertainment and has exhibited remarkable growth in the last few decades bringing about a huge revenue for its stakeholders. Making sure that the revenue generated by a film reaches above the cost of making the movie has always been a prime concern for its investors. A system that estimates the box office return of a movie can be a useful tool for the stakeholders in making informed financial decisions and adjusting the marketing strategy to increase the probability of success. They used Support Vector Machine to classify the user comments between positive and negative classes. Then predicted the success of movies based on the number of positive and negative reviews. Other architectures were also used in different research.

Apart from these factors, there may be other factors such as production house, director, social reach of the cast, preference of the audience, duration, number of users for reviews, language, country, content rating, budget, title, year and release date that can affect the profit generated by a movie. In this literature, we tried focusing on these factors as data related to many of these factors are available before the release of a movie. We wanted to explore how these features contribute to the final revenue that is generated by a film. Also one of our main goals was to be able to predict financial success in the very early stages.

Most of the features mentioned above, for example, the production house, social networking profile of actor/actress, director, budget, genre etc., can be obtained even before the release of movie trailers. So, an attempt has been made in this research to develop a system that will be able to provide us with enough insights about the movie’s performance in the box office by using the parameters described above. Considering the pre-release data provides manifold advantages. The marketing and advertising strategies can be adjusted based on the prediction by the system. Again, since the data related to cast, director, plot, social reach of the actors are available right after making the decision, the early prediction can help in adjusting in these parameters to increase the probability of success in the box office.

In the growing film industry, it is essential to plan ahead and strategies to make a movie successful. It is not always easy to predict the success of a movie. The Box Office Prediction project aims to predict the worldwide box office revenue of an upcoming movie based on data such as its cast, director, crew, promotion, release date, production company and other factors.

Investing in companies from the film industry on the stock market is harder than investing other companies in traditional industries because these film-related companies’ stocks are strongly affected by the financial performance of their upcoming movies in the short term. Hence, accurate and reliable prediction of the Box Office Revenue (BOR) of a movie before releasing is an in pressing need, as it greatly informs the tough investment decision-making process when facing such a high-risk and high-yield opportunity. In addition, the prediction is highly important for advertisement companies that seek to embed their ads in popular movies. Such a prediction can also assist cinemas in scheduling movies and help people choose movies to watch.

# PROPOSED WORK

**Use case diagram**

Diagram

Description automatically generated

**Class diagram**

Diagram

Description automatically generated

#### Data Flow diagram

#### Diagram Description automatically generated

#### Sequence diagram-Profit analysis

#### Diagram Description automatically generated

#### Sequence diagram-Compare results

#### Timeline Description automatically generated

#### Sequence diagram-upcoming movie production

#### Timeline Description automatically generated with low confidence

#### Sequence diagram-Custom movie prediction

#### Timeline Description automatically generated

#### DATASET

#### 4803 rows, 20 columns

#### Features:

#### Budget - The budget includes all costs relating to the development, production, and post-production of a film.

#### Genre - Film genres are categories that define a movie based on its narrative elements.

#### Homepage - Movie Website

#### Id - It is identifier for a movie at The Movie Database. Each movie has an unique identifier.

#### Keywords - Plot Keywords

#### Original Language - Original language is the language of production.

#### Original Title - Original Movie Title

#### Overview - A movie overview gives the reader an idea of the movie.

#### Production companies - A production company, production house, production studio, or a production team is a business that provides the physical basis for works in the fields of performing film.

#### Production countries - Country of origin for movie

#### Release date - Movie release (or premiere) date.

#### Revenue - A film's actual revenue will be determined over time and will include factors like merchandising, product placement fees, television rights and DVD sales.

#### Runtime - Movie Duration

#### Spoken languages - Language of production too and additional languages in case of multiple languages in the movie.

#### Status - This column tells whether the movie is released or not

#### Tagline It speaks directly to the plot of the film. Movie taglines serve to clarify the movie itself.

#### Title - Titles are your audience’s first impression of what your screenplay is about. The title you pick can set the tone for your screenplay and give people a reason to see it

#### USECASES

#### Custom Revenue Prediction

Users can also enter their own parameters for an imaginary movie to test the algorithm and view the results. This can allow researchers to predict the box office revenue for an upcoming movie or a movie still in production.

Traditionally, analysts make predictions based on easily accessible movie metadata, such as genres, budgets, or by referring to market performances of similar movies in the past.Although critic reviews and blog contents are abundantly available, they were not fully exploited until recently with the advances in natural language processing techniques.

Sentiment analysis techniques then became an integral part of movie revenue prediction, and it is motivated the massive amount of user-generated data readily available on the social media.

#### Upcoming revenue prediction

Analyses the revenue of an upcoming movie. Recently, there is an increase in the use of machine learning techniques for movie revenue prediction. The general approach involves predicting the box-office collection of a movie. Box-office is an important determinant of a movie revenue, and it is the total revenue acquired from the sales of movie tickets. Prediction of box-office collection is achieved using mainly supervised learning approaches. Key movie variables like cast or budget are used to train supervised learning algorithms.

Predicting movie revenue has been a research focus for many years. An approximate of 500 movies are produced yearly in the US, with each having an average cost of sixty million dollars. Despite this huge capital investment, the success or profitability of a movie is mostly uncertain. This problem is also found in Chinese, Korean and Indian movie industries. The uncertainties of movie revenues can be associated to the production and distribution of movies, ranging from the actors, directors, and budgets before the release and word-of-mouth marketing, and screen arrangement after the release. Thus, designing effective models for estimating the success of upcoming movies is of paramount importance.

#### Compare results

#### System compares the predicted results with the original results. The data frame returns the actual revenue and predicted revenue.

#### 

#### Profit Analysis

Everyone loves a good movie! From animated classics to franchise action movies, movies appeal to many different groups of people and offer wonderful escapes from reality. The entertainment industry, American produced films, and television shows, is a large part of our nation’s economy, supporting around two million jobs in the country in 2016. The industry feeds a large amount of money into other parts of the economy, for example it paid out around $50 million to local businesses in 2016. The film industry alone generates a massive amount of revenue, with an estimated generated revenue of almost $43 billion in 2017. Since the first movies were produced, people have been trying to figure out what exactly makes a movie a “success”, what common factors. Now, with the growing field of business analytics, it is becoming easier to narrow down some common factors in successful movies. As the demand for new, original content grows, in part due to the rise of streaming platforms such as Netflix, Hulu, Amazon Prime, Disney+, and others producing original movies and television shows, so too does the need for comprehensive analytics about what makes a movie successful. With the ever-growing demand for new movies and the wide range of movie genres, analytics about the performance of movies can give useful information to studios so that they can make the most strategic decisions regarding production and financing. It Studies previous data to examine which factors make a movie successful.

# PROPOSED WORK

## Architecture and Technology Used:

### Software Requirements

Software requirements deals with defining the software resource requirements and prerequisites that need to be installed on a computer to provide optimal functioning of an application. These preconditions are generally not included in the software package and need to be installed separately.

In order to use Box Office Prediction, one should have the following software requirements

* + Operating System: Windows 7 and above
  + Python 3.0 or above
  + Anaconda Navigator
  + pandas, numpy, sklearn, seaborn, matplotlib, streamlit libraries
  + Internet Browser

### Hardware Requirements

Hardware requirements refer to the common set of requirements defined by any operating system or software application and are usually the physical computer resources. In this we look into the architecture, processing power, memory, secondary memory, display adapter and peripherals.

In order to use Box Office Prediction, one should have the following hardware requirements:

* Processor: Intel Core i5 and above
* Memory: 4GB RAM and above

# PROPOSED WORK

## Implementation:

This Project makes use of the Gradient Boosting Regression ML Algorithm for predicting the Box Office revenue. It is available in the sklearn Python library.

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor’s error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels.

The ensemble consists of N trees. Tree1 is trained using the feature matrix X and the labels y. The predictions labelled y1(hat) are used to determine the training set residual errors r1.

Tree2 is then trained using the feature matrix X and the residual errors r1 of Tree1 as labels. The predicted results r1(hat) are then used to determine the residual r2. The process is repeated until all the N trees forming the ensemble are trained.

Random Forest Algorithm was also considered but better results were found using the Gradient Boosting Algorithm since it had a better training score.

# RESULTS

# A screenshot of a computer Description automatically generated

# A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedGraphical user interface, application, Teams Description automatically generatedGraphical user interface Description automatically generatedA screenshot of a computer Description automatically generatedGraphical user interface Description automatically generatedA screenshot of a computer Description automatically generated with medium confidenceA screenshot of a video game Description automatically generatedA screenshot of a computer Description automatically generated with medium confidenceA screenshot of a video game Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a video game Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated with medium confidenceA screenshot of a computer Description automatically generated with medium confidenceGraphical user interface Description automatically generatedGraphical user interface Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated with medium confidenceA screenshot of a computer Description automatically generated with medium confidenceGraphical user interface, application Description automatically generatedA screenshot of a computer Description automatically generated with medium confidence

**CONCLUSION**

Thus, we could comprehend that the prediction of movie success is certainly possible with high percentages of accuracy. So, by using our prediction of our application, production house can evaluate before its 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 casts 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.

# FUTURE WORK

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.

One of the key observations in this preliminary work is that both our neural network approach and the GBR approach have difficulty in classifying the flop movies correctly. We like to include more movies from various parts of the IMDB movie list in our future work and not only the top 100 movies as we did in this work. We believe that will increase the classification accuracy of the flop movies. We also like to add more classes to categorize movies such as ‘bomb’, ‘good’, and ‘flop’ and more input data attributes about movie content such as the type of movie and visual effect.

We also intend to give users their own accounts and store their created movies and results for further use.

# REFERENCES

Kaggle. TMDB. Retrieved from <https://www.kaggle.com/tmdb/tmdb-movie-metadata>

Donges, N. (2018, Feb 22). The Random Forest Algorithm. Retrieved from <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>

Jha, V. (2017, June 18). Decision Tree Algorithm for a predictive model. Retrieved from <https://www.techleer.com/articles/120-decision-tree-algorithm-for-a-predictive-model/>

Microsoft Azure. (2017, Nov 03). The Team Data Science Process Lifecycle. Retrieved from <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-scienceprocess/lifecycle>

Statistics Solution. (2013). What is Linear Regression. Retrieved from <https://www.statisticssolutions.com/what-is-linear-regression/>

Machine Learning Mastery. Introduction to Gradient Boosting Regression. Retrieved from [https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-](https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/) [learning/](https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/)

scikit-learn. Gradient Boosting Regressor. Retrieved from

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html>

Streamlit. Streamlit Documentation. Retrieved from <https://docs.streamlit.io/>

The Movie Database. TMDB API. Retrieved from <https://developers.themoviedb.org/3>