iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 Issue: VI Month of publication: June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44113

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Movie Data Analysis and Prediction

G. Pranay Mukund[1], Kadari Naveen Kumar[2], Aruna Kumari Kumbhagiri[3]
[1, 2]Department of Electronics and Computers Engineering, Sreenidhi Institute of Scienceand Technology, Hyderabad, India
[3]Assistant Professor Department of Electronics and Computers Engineering, Sreenidhi Institute of Scienceand Technology, Hyderabad, India

Abstract: With the use of Big Data in CinemaAnalysis, we are able to assess the model with better precision and eliminate the speculation that typically accompanies theprocess. The primary goal of this research is to study and create a training data that can forecast the movie's income. We used Kaggle's information, which includesinformation on 3000 films, includinginformation on the film's title, cast, and budget. Two separate classification methods areused in this project to evaluate, visualise,and train the collection. Regularization and Strange Wooded are the two methods.Algorithms are evaluated based on their RMSE values, and the one with the lowest score is selected. Our last prediction was for film in the collection that did not have a revenue associated with them.

I. INTRODUCTION

The term "Big Data" represents a set of documentation that is both large and growing at an exponential rate. Conventional data processing technologiesare unable to store or handle informationeffectively because it is so massive and complicated. It's a lot of data, but it's also a lot of data. Demand forecasting and other machine learning programs also make use of this technology. Judgement call in the movieindustry rely on big data techniques to help respective film firms succeed in a crowded marketplace. With the help of this information, they are able to establish realistic objectives and learn how to achieve them.

II. LITERATURE SURVEY

We have done a lot of literature review on the similar movie revenue prediction projects. We have got some of the existing projects.

Title of the paper 1: Early prediction of movie Box office success. Based on Wikipedia Activity Big Data:
Description: They presented the results of developing a simple statistical model for movie financial performance based on internet user's cumulative activity data.
By calculating and evaluating the activity level of editors and viewers of the corresponding entry to the movie in Wikipedia, the well – known online encyclopedia, they demonstrated that the success of a movie can be predicted much before its publication.

Title of the paper 2: Sentiment Analysis of Movie Reviews using Machine Learning Techniques.
Description: It is the analysis which made based on emotions and opinions of any form. Sentiment analysis is also named as opinion mining. This type of methos is useful when we given a content to a particular person as a source to know the sentiment. It is useful to explain the view of a bunch of people or a person. In this sentiment analysis they used techniques like Naïve Bayes, K-Nearest Neighbour and Random Forest.

Title of the paper 3: Movie Success prediction using Data Mining
Description: In this model to predict the success and failure of a movie they used a mathematical model based on some attributes. Some of the attributes used for predicting is genre, director, and budget. To dig out the patterns and trends which will be useful in predicting movie success they used data mining process and applied to movie database. In this model also they used data cleaning and integration process.

Fig 1: Big Data on Movies

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Ridge regression and the Randomised forest technique are being used to build a classification model in this project. Main objective is to develop a hardware modelthat can predict overall sales of a latest film based on available resources, playable demos and classifications and production companies and nations. Kaggle has providedus with a sample of 3000 movies (from 1960to 2017) with all the important facts on everyone one of them (like cast, crew, budget, popularity, date, Genre etc..). It reduces the financial burden on film producers who plan to make a movie.

III. METHODOLOGY

Three stages are included in this operation. Among the three stages are Pre-processing, Modeling, and Testing. Each phase has its own internal procedure. This project explains each step in detail. They're as follows:

A. Pre-processing

In order to develop a model, dataset is an essential part of the process. Data collecting, verification, analysis techniques, and manipulating categorical values are just afew of the phases involved in this process. Model adoption and assessment benefit from increased data quality. This stage consists ofthe following steps:

B. Data Collection

The collection of data Kaggle, a cinematic database, provided the data for this dataset, which spans the years 1960 to 2017. It includes details such as the cast, crew, prequel, popularity, budget, and genre of the film. " The picture of the database is shown below.

```
data = pd.read_csv("/content/drive/MyDrive/BDA project/train.csv")
print(data.shape)
data.head(n=2)
```

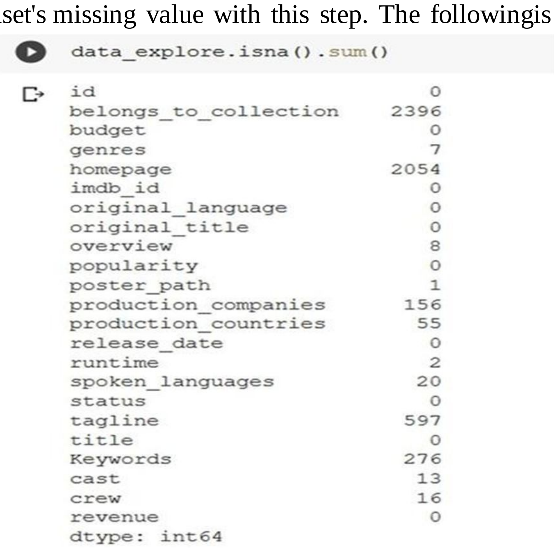
(3000, 23)

	id	belongs_to_collection	budget	genres	homepage	imdb_id	original_language	original_title	overview	popularity	
0	1	[{'id': 313576, 'name': 'Hot Tub Time Machine ...	14000000	[{'id': 35, 'name': 'Comedy'}]	NaN	tt2637294	en	Hot Tub Time Machine 2	When Lou, who has become the "father of the In...	6.575393	/tQtWuw
1	2	[{'id': 107674, 'name': 'The Princess Diaries ...	40000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	NaN	tt0368933	en	The Princess Diaries 2: Royal Engagement	Mia Thermopolis is now a college graduate and ...	8.248895	/w9Z7A0G

Fig 1: Dataset of 3000 movies

C. Data Cleaning

We acquired raw data in the form of a dataset. Everything from the cast to the crew to the movie's success to its genre and subgenres are included. Data preparation is acritical stage in the process of transforming raw data into data that can be used to train models. We are deleting all of the dataset's missing value with this step. The followingis an example of what I'm talking



```
data_explore.isna().sum()

id                         0
belongs_to_collection   2396
budget                     0
genres                     7
homepage                2054
imdb_id                    0
original_language          0
original_title             0
overview                   8
popularity                 0
poster_path                1
production_companies     156
production_countries      55
release_date               0
runtime                    2
spoken_languages          20
status                     0
tagline                  597
title                      0
Keywords                 276
cast                      13
crew                      16
revenue                    0
dtype: int64
```

Fig 3: Checking Null values

Figure 3 displays the dataset set's data type in a separate cell. There are 2034 empty values in the fig-3 main page category.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

D. Data Analysis

The third phase in this procedure is dataanalysis. In order to carry out the next stages, it is critical that the data be understood. This stage includes data visualisation. Using a graphical illustration of ordinal data data helps us understand the data better and may lead us to the next step in the process.

Fig-4 shows the top 20 highly interestingfilms, with the X-axis indicating popularity and the Y-axis indicating the description of the movies. Wonder Women, with a highapproval rating of 294, is the most well-liked film of all time.

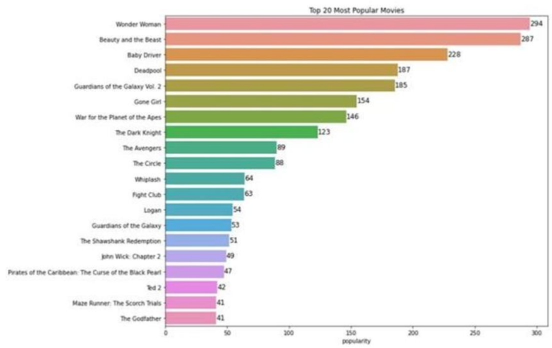


Fig 4: Graphical representation of TOP 20 popularity movies

Fig-5 shows the top 20 highest-grossing pictures, with the X-axis indicating money (as a million dollars) and the Y-axis indicating the name of something like the movies. The Avengers, with a gross of $1,519 million, is the highest-grossing film.

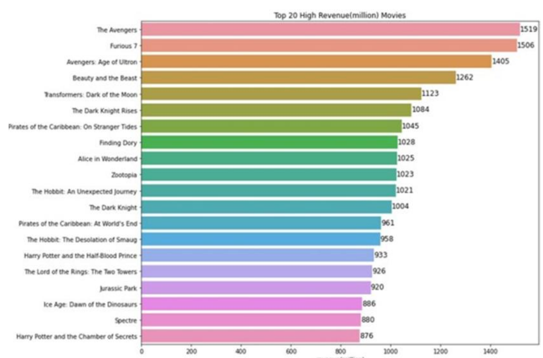


Fig 5: Graphical representation of TOP 20 high revenue movies.

Below is Fig-6, which displays data on the top 20 highest-grossing blockbusters, withprofit (in millions of dollars) plotted on the X-axis and movie titles as seen on the Y- axis. Pirates of something like the Caribbean: On Random person Tides had a budget of 380 million dollars, making it the most expensive film ever made.

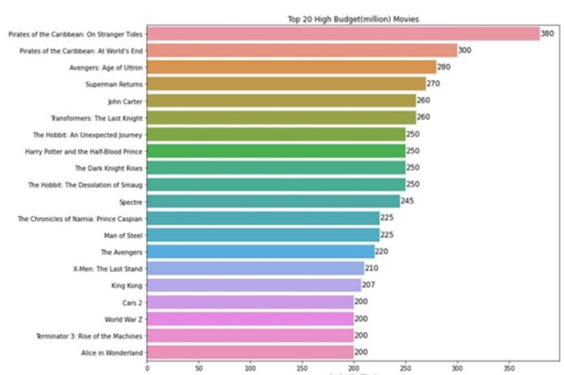


Fig 6: Graphical representation of TOP 20 high budget(million) movies

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

With budget(in millions) on the X-axis as well as movie title on the Y-axis, Fig. 7shows the top 20 biggest producing films of all time.

1316 million dollars is the highest grossing film of all time, making it the most profitable film of all time.

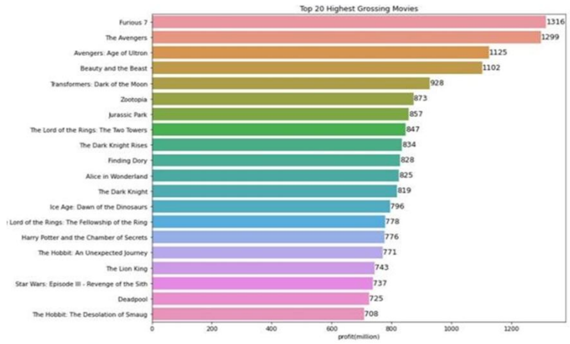


Fig 7: Graphical representation of TOP 20 highest Grossing(million)

The number of films in various genres isdepicted in Fig. 8 below. The X-axis shows the various genres, while the Y-axis shows the total number of films in each category. According to the graph below, there have been 1531 films based mostly on genre of drama produced in theatres.

In Fig-9, the link between Musical stylesand Mean average Popularity is depicted.The X-axis represents genre kinds, while theY-axis depicts the level of popularity foreach genre.

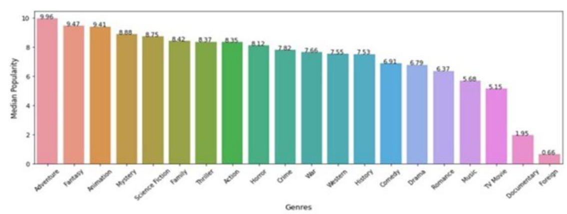


Fig 9: Genres versus Median popularity

Fig-10 shows us the total amount of money spent and the amount of money made for a certain genre. When it comes to the X-Axis we used genres, and the Y-Axis was used to add up the total number of genres (in million).

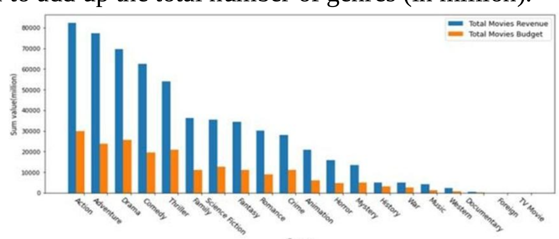


Fig 10: Budget and revenue of particular Genre

Fig. 11 illustrates the revenue-to-budget connection. X-axis was the budget, and Y- axis was the revenue generated by the company.

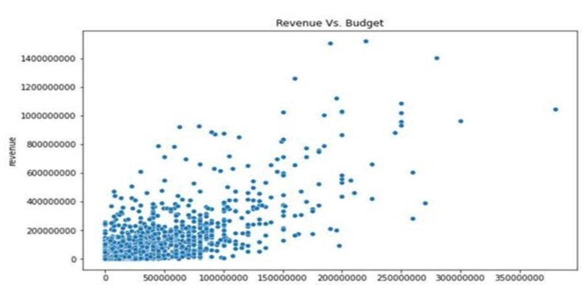


Fig 11: Scatter plot representation of Budget vs Revenue.

The link among both revenue and popularitymay be seen in the graph below (Fig. 12). Attraction and profitability were measured on the X-axis and the Y-axis.

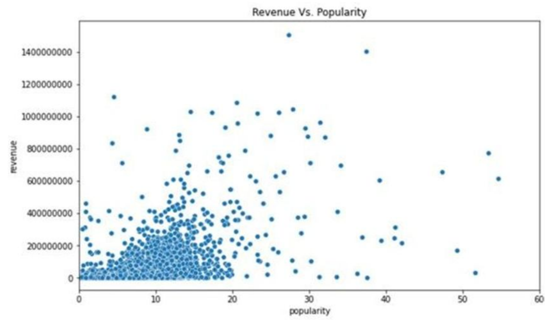


Fig 12: Scatter plot representation of Revenue vs Popularity

Revenue and movie runtime are shown in the following Fig-13. Runtime was plotted on the X-axis, while revenue was plotted on the Y-axis.

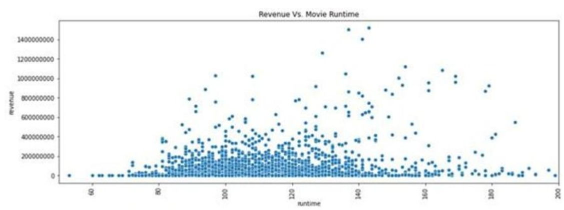


Fig 13: Scatter plot representation of Revenue vs Movie Runtime.

The movie's total revenue for each of itsrelease years is depicted in the graph below.
In the following graph, movie ticket sales for each of the months in which the films were released are shown.

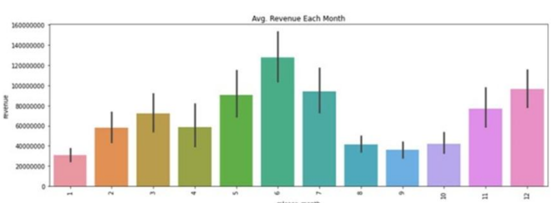


Fig 15: Scatter plot representation of Revenue vs Movie Release month.

An overview of worldwide box office receipts for various languages is shown in the table below.

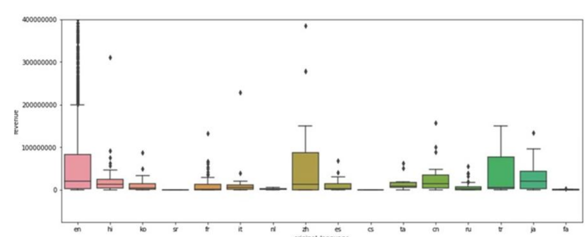


Fig 16: Scatterplot representation of Revenue vs Movie Original language.

We've also figured out how each feature relates to the others, and we expressed that information graphically. There are pleasant and unpleasant correlations between thevariables we observed.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

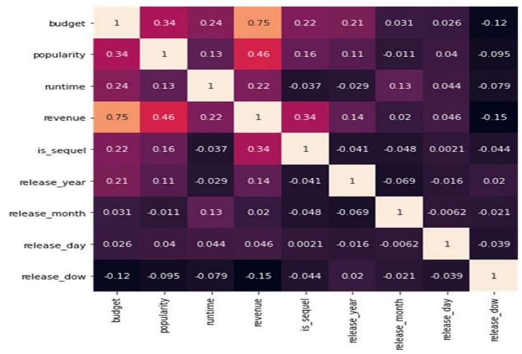


Fig 17: Correlation of each feature of dataset

E. Handling Categorical Values

Attribute data and category values are incompatible with neural network models. So, the categories values must be handled before the model can be implemented. Handles is just like the log - log plot of the values. It aids in the quantitative transformation of attribute variables.

Comedy	Drama	Family	Romance	Thriller	Action	Animation	Adventure	Horror
1	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0
0	1	0	0	0	0	0	0	0

Fig 18: Dataset and datatypes after applying one hot encoding

IV. RESULT AND DISCUSSION

Several films with unknown box office grosses were fed into the model once it had been trained successfully. Even though there were a few null values and similar factors like genres, size, and appeal, his model was able to accurately calculate the gross receipts for all 4000 movies included in the data set. Titles like Pokemon: The Rise of Darkrai, Attack of the 50 Walking Woman, Love, Incedies, and Inside Deep Throat are expected to bring in a lot of money. Output.csv contains a total of 4000 films' estimated revenues, which have not been tied to the profit.

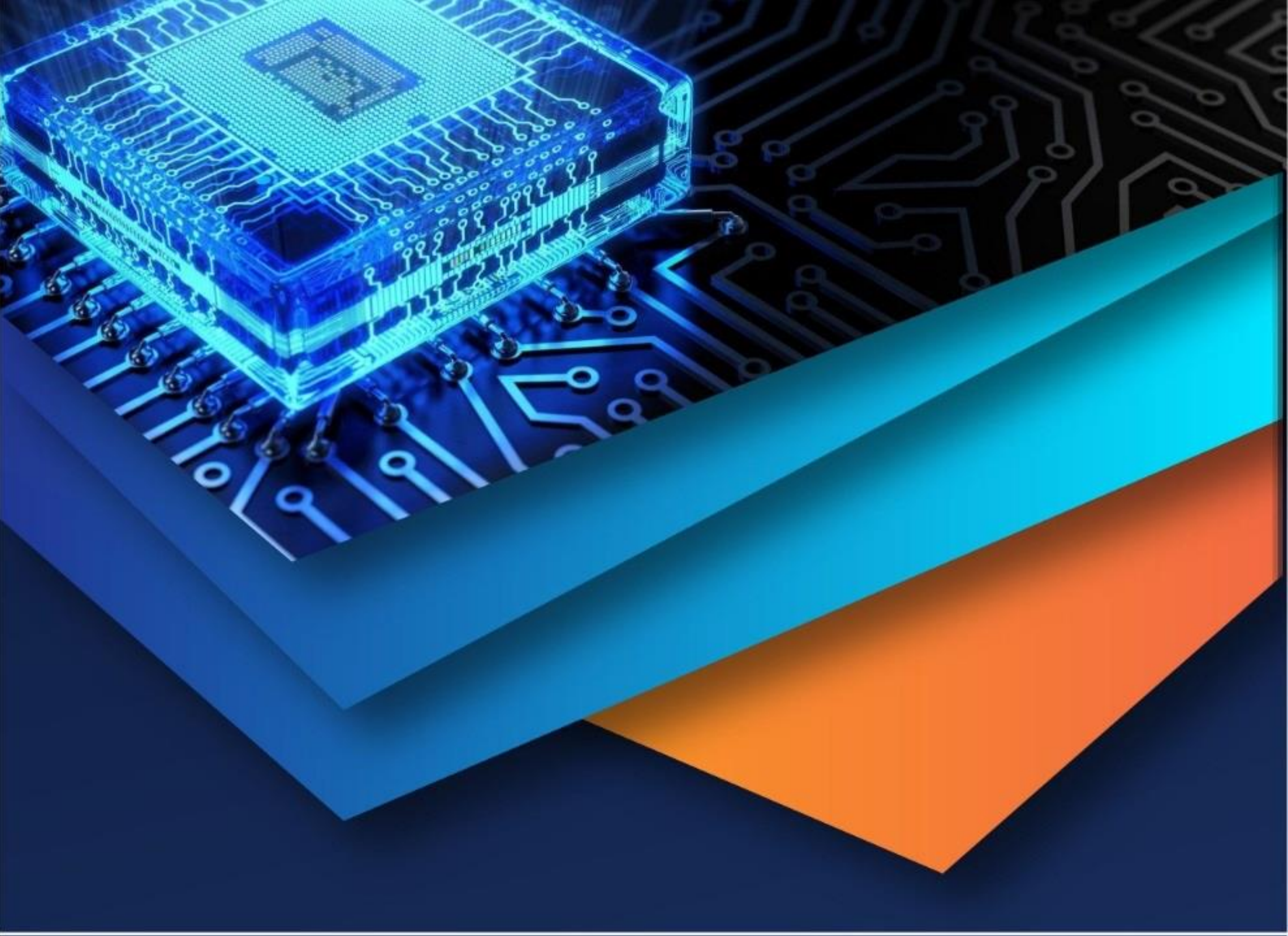
V. CONCLUSION

We can now estimate the model's income more accurately thanks to Big Data in Cinema Analysis, reducing the uncertainty that sometimes accompanies this type of analysis. The primary goal of this research isto study and create a classification modelthat can forecast the movie's income. We used Kaggle's dataset, which includes information on 3000 films, including information mostly on film's title, cast, and production. Two separate classification methods areused in this project to evaluate, visualise,and train the dataset. Regularization and Randomised Forest are the two methods. Algorithms are evaluated based on their RMSE values, and the one with the lowest score is selected. Last but not least, we've estimated the box office receipts for 4000 pictures in the database that weren'tpreviously associated with their box office receipts.

	title	revenue
0	Pokémon: The Rise of Darkrai	4.312409e+06
1	Attack of the 50 Foot Woman	1.574562e+06
2	Addicted to Love	6.327415e+06
3	Incendies	1.014175e+06
4	Inside Deep Throat	6.030557e+05

REFERENCES

[1] Early Prediction of movie Box-Office success. Based on Wikipedia Activity Big Data. Marton Mestyan, Taha Yasseri, Janos Kertesz, 2012.
[2] Sentiment Analysis of Movie Reviews using Machine Learning Techniques. Palak Baid, Apoorva Gupta, Neelam chaplot, 2017.
[3] Movie Success Prediction using Data Mining. Anantharaman V, Ebin G. Job, Neha sam, Sheryl Maria Sebastian, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY