Employability of the Data Analysis of the Relevant Dataset Based on Classification and Clustering Algorithms in the Effective Prediction of Movies

Saniya Malik

DAV Police Public School, Gurugram

ABSTRACT

In Movie Analysis, Big Data enables us to evaluate the model more precisely and eliminate process-associated speculation. The main purpose of this research is to investigate and produce training data that can predict the movie's earnings. We used the data from Kaggle, which included information on 3,000 films, including the title, cast, and budget. The collection is evaluated, visualized, and trained with the help of two classification techniques in this project. The two methods are Regularization and Strange Wooded. The algorithm withthe lowest score is chosen after being evaluated using its RMSE values. Our most recent prediction was for a movie in the collection that didn't bring in any money.

INTRODUCTION

The term "Big Data" refers to a large and rapidly expanding collection of documentation. It isa lot of data all at once. Conventional data processing technologies require assistance to effectively store or handle it because of their size and complexity. Demand forecasting and other machine learning programs also make use of this technology.

In the film industry, judgment call utilizes big data strategies to support each film company's success in a competitive market.

They can learn how to set and achieve realistic goals with this information.

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 encyclopaedia, 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 give 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 likeNaïve 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.

METHODOLOGY

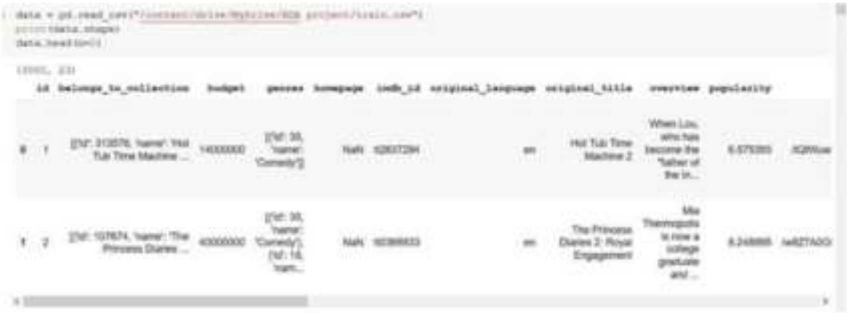
This operation is carried out in three stages. Pre-processing, modelling, and testing are the three stages. There are internal procedures for each phase. Each step is explained in detail in this project. These are some of them:

Pre-processing

The dataset is a crucial component of the model-making procedure. A few stages in this process include data collection, verification, techniques for analysis, and manipulating categorical values. The steps that make up this stage are as follows:

Data Collection

The cinematic database Kaggle provided the data for this dataset, which spans the years 1960 to 2017.It includes information about the cast, crew, Popularity, budget, and film genre. The database's image can be seen below.



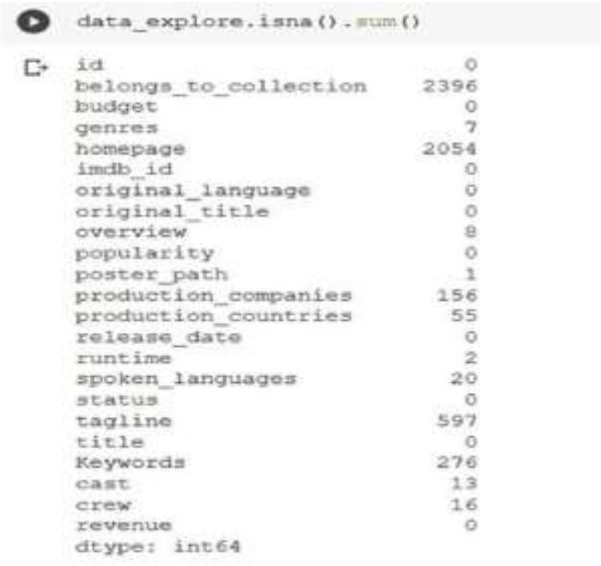
```
data = pd.read_csv("/content/drive/MyDrive/ML project/train.csv")
print(data.shape)
data.head(n=2)

(3000, 23)
   id belongs_to_collection              budget    genres                      homepage imdb_id   original_language original_title                          overview                                              popularity
0  1  [{'id': 313576, 'name': 'Hot        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
      Tub Time Machine ...
1  2  [{'id': 107674, 'name': 'The        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
      Princess Diaries ...
```

Fig 1: Dataset of 3000 movies

Data Cleaning

It is necessary to properly prepare the data to transform raw data into data that can be used to train models. We obtained a dataset of raw data. The movie's success, cast, crew, genre, and subgenres are included. This step removes all of the missing values from the dataset. An illustration of what I'm talking about can be found below.



```
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

Data Analysis

It is the third step in this procedure. Understanding the data is necessary for completing the subsequent steps. Data visualization is included in this stage. A graphical representation of ordinal data enhances our comprehension and may direct us to the next stage of the process.

The top 20 most intriguing movies are depicted in Fig. 4, with the popularity axis representing Popularity and the movie description axis representing description. WonderWoman is the most popular film of all time, receiving a high approval rating of 294 per cent.

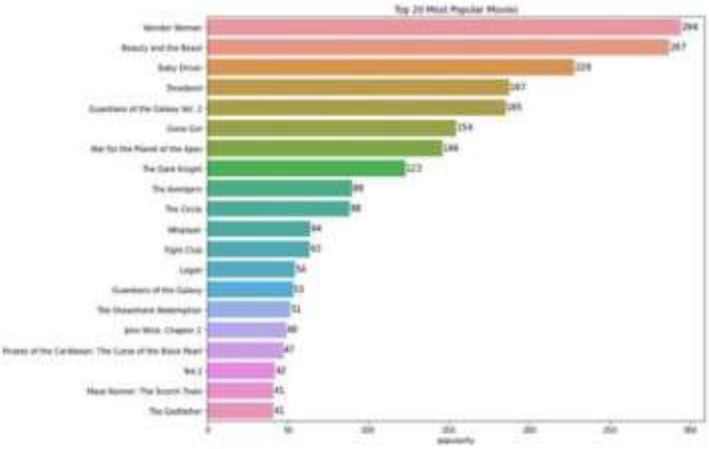


Fig 4: Graphical representation of TOP 20 popularity movies

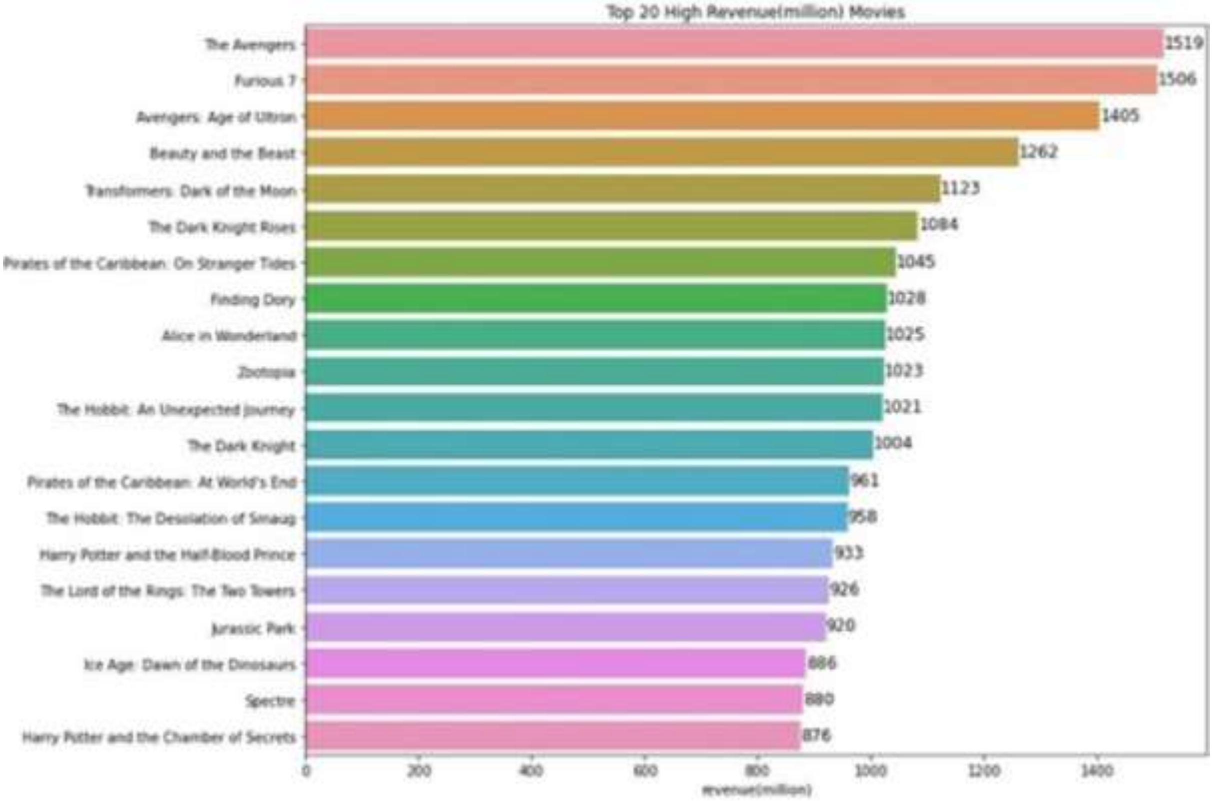


Fig 5: Graphical representation of TOP 20 high revenue movies.

The data for the top 20 blockbusters can be found in Figure 6, which displays movie titles on the Y-axis and profit (in millions of dollars) on the X-axis. Random Person Tides was the most expensive movie ever made, costing 380 million dollars and based on Pirates from the Caribbean.

The film One Thousand Three Hundred and Sixty-Six Million Dollars is the most successful film ever made.

The number of films in various genres below. The various genres are depicted on the X-axis, and the total number of films in each category is depicted on the Y-axis. The graph below shows that 1531 movies, primarily of the drama genre, have been shown in theatres.

The relationship between musical styles and Mean Average Popularity. The genre types are shown on the X-axis, and each genre is shown on the Y-axis.

Handling Categorical Values

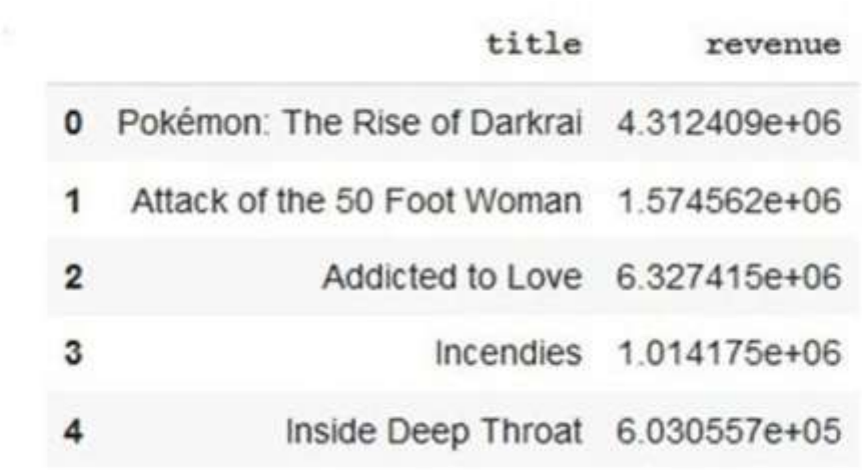
Neural network models cannot work with attribute data or category values. As a result, the values of the categories need to be dealt with before the model can be used. Attribute variables can be quantitatively transformed with its assistance. The handles are analogous to the values' log-log plot.

RESULT AND DISCUSSION

Following the successful training of the model, several movies with unknown box office tickets were fed into it. Despite having a few null values and similar factors like genre, size, and appeal, his model correctly calculated the gross receipts for all 4000 movies in the data set. Pokemon-like films: It is anticipated that Inside Deep Throat, Love, Indies, and Rise of Darkrai will generate a significant amount of revenue. The estimated revenues for 4,000 filmsin Output.csv are unrelated to the profit.

CONCLUSION

We can now accurately estimate the model's gain thanks to Big Data in movie analysis, whichreduces the uncertainty sometimes associated with this kind of analysis. Creating a classification model that can accurately predict the movie's earnings is the study's primary objective. We used the dataset from Kaggle, which contains information on 3,000 films, primarily its title, cast, and production. The dataset is evaluated, visualized, and trained using two classification techniques in this project. The two methods are Randomized Forest and Regularization. After examining its RMSE values, the algorithm with the lowest score is selected. Last, we calculated the box office receipts for 4000 movies in the database that had never previously been linked to 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.