Analysing Movie Success Based on Machine Learning Algorithm

Ashutosh Shankhdhar¹, Vinay Agrawal², Vikram Rajpoot³*
¹, ², ³Dept. of Computer Engineering and Applications GLA University, Mathura U.P, India
³*vikram.rajpoot@gla.ac.in

Abstract The Film Industry is producing several hundreds of movies paving the way for the United States for third position in the list of giants of film Industry of the whole world [1]. The expenditure on these movies are of the ten to eleven figures or we can say thousands of millions of dollars, which ensure their box office success, is absolutely essential for the survival of the industry. Predicting what sort of movies are going to earn more and what type is going to come up short before the release, may profit the houses extraordinarily because it will also empower them to work on their promoting efforts which itself need numerous dollars, appropriately[2]. Furthermore, it would likewise assist them with knowing when it is generally suitable to deliver a film by taking a look at the general market. Along these lines, the examination of film achievement is critical to the business. ML (Machine Learning) algorithms [3] calculations broadly need to make expectations like development inside the stock trade, interest for items, nature of tumors, and so on.

Keywords- Data Wrangling, Data Cleaning, Exploratory Data Analysis.

1. Introduction
In the US just, the 2000s of movies are delivered per annum. Since the 1920s, the American Movies have earned additional cash per as compared to the other nation [4]. Films in America might be a multi-billion dollar industry where every single movie acquires over a billion dollars. Certain famed production houses control the maximum part of this business, where billions of dollars are spent in promotion only. Promotion and ads expenses are more intense than the whole spending plan of the movies. Most of the times this investment leads to big loss to the makers [5]. W.B.[Warner Brothers], one of the giants of creation houses, had a drop in their incomes a year ago, despite the swelling and subsequently the expanded number of flicks delivered. In the event that it had been some way or another conceiveable to see heretofore the probability of accomplishment of the newly released films, the assembly houses could change the let out of their movies to increase the greatest benefit. they may utilize these findings to ensure the outcome i.e when the market is profitable and when it's not. It paves the way for a desperate requirement for such programming to be created. A lot of scholars have attempted to achieve this objective of study film incomes. The paper is coordinated as follows. Area 2 portrays the function of information set assortment and pre-processing in the information handling. Execution has been examined in segment III. Results have appeared in segment IV and Section V finishes up the paper.
2. Dataset Collection and Processing

2.1 Dataset Collection
Dataset was initially from Kaggle, and given by Udacity. As indicated by Kaggle's presentation page, the information contains data that is given from The Movie Database (TMDb)[6]. It gathers 5000+ movies’ fundamental movie data and film matrices, having user rating, revenue, and popularity ratings. These measurements are frequently observed as to how fruitful these films may be. The film's fundamental data contained things like the cast, chief, watchwords, runtime, classes, and so on (Comment: The dataset on the Kaggle page might be refreshed to a newer version.) And the entire dataset term covers from 1960 to 2016.

![Data Definition Diagram](image)

Figure 1. Data Definition

2.2 Data Pre-processing
The dataset we get is very mislaid, noisy, and conflicting information because of the big size and different origin, various sources [7]. We generally utilized TMD Dataset. The principal issue with datasets was mislaid fields. To remove this mislaid field issue so we embraced a way that utilizes both median and mean as a focal inclination. Then eliminated duplicated things.

3. Data Integration and Transformation
The data that we got from three unique assets Wikipedia, Rotten Tomatoes, and IMDb[8] were coordinated into one TMDb data set. So that the coordinated information is changed or united so the analysis could be effective and simpler. Dataset is blended in with both nominal and numeric characteristics. We utilized the proportion of central tendency of Box office income to change comparing nominal characteristics over to numerical.
Table 1. Features for characterising the movie attributes

| Types              | Characteristic                                      |
|--------------------|----------------------------------------------------|
| Nominal values     | original_title, cast, director, overview, genres,  |
|                    | release_date, production_companies                 |
| Numerical values   | id, imdb_id, popularity, budget, revenue, runtime,|
|                    | vote_count, vote_average, release_year,            |
|                    | budget_adj, revenue_adj                            |

4. Implementation

4.1 Data Wrangling
Data wrangling is a process of taking and standardizing imperfect and disorganised raw data so that you can easily access and analyze it. Data wrangling, now and then referenced as data munging, is a way which is used for modifying and planning given dataset from one "useless" dataset structure into another form with the aim of making it more appropriate and necessary data form for an assortment of downstream purposes, for example, analytics [6]. It also includes mapping data fields from source to destination. A data wrangler might be the one that carries out these transformation operations.

4.2 Data Cleaning
Data cleaning is the method of updating data to make sure that it is free of errors and incorrect details. Often known as data cleaning, this includes finding incomplete, wrong, irrelevant, missing and dirty sections of the dataset and then removing or cleaning up the dirty parts of the data. Data cleaning is about finding the best approach to amplify an information set's exactness without essentially erasing information [7]. Data can be cleaned by eliminating or changing data that are incomplete, irrelevant, duplicate, incorrect, or inappropriately formatted.

Months with highest number of movie release in all years –

Figure 2. Month vs No. of movies releases
4.3 Data Analysis

Analysis can be considered as more of a process in order of cleaning, and transforming data to get valuable information for industry dynamics. The focus of Analysis of data is to extricate helpful information from data and settle on the choice dependent on the data analysis [8]. Presently something very similar an analyst accomplishes for business purposes named Data Analysis. According to this graph in Figure 2, September and October are those months in which the highest number of movies are released.

Movie length liked by audience -

![Runtime Vs Popularity](image1)

**Figure 3.** Runtime vs Popularity

According to the plot we can say that movies in the range of 120-200 runtime are more popular than other runtime movies. Because it is boring to see the long duration movies.

Month that made the highest average revenue

![Average revenue by month (1960 - 2015)](image2)

**Figure 4.** Revenue collected by month (1960-2015)

According to this graph May, June, November and December are those months in a year in which the highest revenue collected by the movies.
Genres with the highest release of movies –

![Popular Genres](image)

**Figure 5. Popular Genres**

According to this plot Drama (4761) genre has the highest release of movies followed by Comedy (3793) and Thriller (2908).

Correlation of properties associated with movies with high revenue—

![Properties associated with successful movies](image)

**Figure 6. Properties associated with successful movies**

5. Results

Budget vs Revenue: After analysing the data, it was noticed that with increase in budget, the revenue amount increases proportionally with a positive correlation (0.68) This implies there is a decent likelihood that movies with higher investment bring about more revenues.

Profit vs Budget: The research conveyed that with a profit growth budget goes in hands with profit amount with a positive correlation (0.53) This implies there is a decent possibility that movies with higher investment bring about more prominent Profit.
Release Year vs Vote Average: After investigating the findings this can be concluded that release year and voting does not affect the movie success and the same is conveyed by negative correlation (-0.11). This implies that movie ratings don't rely upon the delivery year.

Popularity vs Profit: When the outcome was analysed then it was perceived that as the newly released movie gains popularity the profit comes as a bonus providing a positive correlation (0.61). It means that movies with high popularity tend to earn high profits.

Popularity vs Revenue: In this research it was detected that popularity contributes directly towards revenue as is shown by the positive correlation (0.68). Hence it can be derived that if the popularity of movies is high then the revenue too is going to rise up.

Vote Average vs Revenue: The correlation between revenue and vote average is 0.2069. So vote average is not highly related to the revenue.

Runtime vs Revenue: The correlation between revenue and runtime is 0.2378. So runtime is not highly related to the revenue.

6. Conclusion and Future Work

After analysing the whole dataset a lot of interesting findings came into light, which contains such plentiful information that it can dig out the properties of successful movies as well as different kinds of matrices so that results can be cross-compared. A basic data analysis process was performed using the tactics learnt so far.

There were four popular genres: drama, action, comedy and thriller and Drama came out to be the most popular genre followed by action, comedy and thriller. 2014 was the year in which production of movies was at its peak. The audience prefers short movies over lengthy movies, where the average runtime of movies is decreasing each year. May, June, i.e. summer vacation and November, and December winter vacation are the months when movies earn large profit, high popularity proving to be the most suitable month for releasing to ensure maximum earnings. Thus if a production house wants to earn more profit then these findings must be taken into consideration strictly. High revenue paves the way for greater budget. Warner Bros, Universal Pictures and Paramount Pictures production companies begged the top positions and have more lifetime profit than other production companies. Movies with higher budgets have shown a corresponding increase in revenues.

In future this research is going to further advance towards an application or an app which will generate a report after analysing the previous trends i.e by taking into consideration information like what genre was liked most and what length of movie was preferred. Was it in controversy etc. In the app the producer will fill all the info like budget, releasing month, genre, and important factors and based on those factors the report will predict whether the movie is a success or not.

7. Limitations

It's not a 100 percent guaranteed solution that this formula is going to work, but it shows us that we have a high probability of making high profits if we had similar characteristics as such. If we release a movie with these characteristics, it gives people high expectations from this movie. This was just one example of an influential factor that would lead to different results, many have to be taken care of. During the data cleaning process, I split the data separated by '|' within records for simple parsing during the investigation stage. This expands the time taken in calculating the outcome.

References

[1] Ahmad, J., Duraisamy, P., Yousef, A., & Buckles, B. (2017, July). Movie success prediction using data mining. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-4). IEEE.

[2] Joshi, A. M., & Hanssens, D. M. (2009). Movie advertising and the stock market valuation of studios: a
ase of “great expectations?” *Marketing Science*, 28(2), 239-250.

[3] Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd.

[4] Sklar, R. (2012). *Movie-made America: A cultural history of American movies*. Vintage.

[5] Simonoff, J. S., & Sparrow, I. R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3), 15-24.

[6] Kaggle datasets. (https://www.kaggle.com/tmdb/tmdb-movie-metadata)

[7] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining Concepts and Techniques*, 3rd ed. MA: Elsevier, 2011, pp. 83-117

[8] The International Movie Database (IMDb). https://www.imdb.com/

[9] Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N. H., ... & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization, 10*(4), 271-288.

[10] Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning* (Vol. 479). John Wiley & Sons.

[11] Brandt, S. (1998). *Data analysis*. Springer,