Revenue Prediction Model

Abhishek Verma
New York University
New York, NY, USA
av2783@nyu.edu

Ishita Jaisia
New York University
New York, NY, USA
ij2056@nyu.edu

FNU Shivanshi
New York University
New York, NY, USA
ss14396@nyu.edu

***Abstract*—The film industry is the major contributor to the world-wide economy. Every year more than hundreds to thousands of movies are being released . Hence, analyzing the box office success of the movie can be crucial. This paper focuses on analysing the factors that determine the movie performance in terms of profit and popularity. We use map-reduce to learn insightful characteristics of the data and use Hive to gain analytical understanding. For statistical observation, analysed data is plotted in graphs. We also introduce a Return on Investment(RoI) prediction model to determine the success of the movie.**

***Index Terms*—movie, analytics, map-reduce, big data, hive, hadoop, prediction model**

I. Introduction

Movies have become a remarkably effective medium for entertainment for people over the past few decades. Among numerous movies that have been released over the past decade some of them have went ahead and gained high profits. The film industry has become high profile with a highly variable revenue stream. This billion dollar industry keeps growing as the demand for new and original content grows due to rise of the streaming platforms such as Netflix, Amazon Prime Video, Hulu, HBOMax, Disney+ and others which produce original movies. Revenue generated by the movie industry reached $99.7 billion in 2021 including global theatrical and home entertainment. People have been trying to figure out what make a movie successful and what are the common factors that affect it, since the first movies were produced. There are a number of factors on which a movie's success depends like the cast, crew, director, producer, release date, plot, budget, duration of the movie and the reactions of the viewers. As the business analytics field grows it has become easier to find a relationship between the "succes" of the movies and the factors common to them [1] [2].

We try to determine how accurately can we predict the revenue of the movie prior to its release. Are there certain genres of movies that make them more likely to be moneymakers? Does the budget of a movie play a big role in its success? Is there a correlation between a director winning an award and their popularity and profit? Predicting the revenue of a movie is important due to the risks involved despite the high investment involved in its production [3].

II. Motivation

With the ever growing demand for new content in movies, analytics about the movies could help us understand more about the performance of the movie on the box office. The number of exogenous variables that affect a movie's performance are huge which make the revenue prediction difficult. However, we are in the era of data science where high volumes of data can be efficiently processed and modeled according to our requirements. This analytics can help studios to make decision strategically about financing, the production stages and the distribution stages and determine what factors will help them increase the profits for their movies.

III. Literature Survey

In [4] the authors have predicted movie grosses. In this paper, they have mainly focused on US domestic gross. They also examine foreign gross but not to a very large extent. The analysis done by the authors is based on new movies released in the USA in the calendar year 1998 from the Internet Movie Database. They first predict the movie grosses based on information that is available before release. They made use of MPAA rating, Genre, Number of Best Actors, Number of top dollar actors, and whether the movie was a summer release or not as variables in the model, whether or not the movie is a sequel to an earlier movie. The authors claim that apart from the aforementioned predictor variables other potential predictors did not play a very crucial role in the prediction model. The authors state that the opening weekend of a movie's release accounts for 25 percent of the total box office gross. Expecting the opening weekend grosses would be highly predictive of gross but this doesn't take into account the different movie release patterns. Since some movies will open at once on thousand screens and others won't. The authors used the aforementioned knowledge of opening weekend knowledge to build a better prediction model. The results proved that the opening weekend grosses do help to predict gross revenue more accurately than other models.

IV. Proposed Idea

We collect data from Kaggle. The data we used is a total of three datasets all of which are static in nature. We clean our data and pre-process it using MapReduce jobs. We filtered out the data and acquired the necessary information. Using this information we profile our data and and learn more about it to come up with the answer to the following research questions.

- Is there a relationship between the release month of the year and the movie's profitability and popularity? If yes,

which months as release months turn out to be most lucky for movies in terms of popularity and profit?
- How has the time duration been affecting high Profits, High Voting Average and High Popularity over the years from 2007 to 2017?
- Does the popularity of a director in a particular calendar year affect the chances of them being nominated for an award at the Emmy Awards?

We also create a RoI prediction model which helps us understand if we can predict the success of movies based on a couple of factors.

V. Data Sources

We use three different data sets for our analysis. All the datasets were collected from Kaggle. We selected these datasets because they contain information about different entities. The datasets are discussed in detail below.

1) The Movies Dataset:
   This dataset is an ensemble of data collected from TMDB and GroupLens. This is a database of 45000 movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages [5]. The files in this dataset that are useful to us are:
   - movies_metadata.csv: This has details about the genre, budget, language, popularity, etc. There are a total of 45573 number of records in this file out of which 186 have missing columns. Budget, Genres, Movie ID, Popularity, Revenue, Title and VoteAvg have 45386 values which are not NULL. Movie ID has 30 and Title has 3175 duplicate entries. There were 84 entries in Release Date which had NULL value.
   - credits.csv: This contains the crew and cast information of the movies. There are 43032, 44678, 45444 unique values of cast, crew and movie id respectively. The total number of records is 45477 and none of the above mentioned attributes had a NULL value.

   There are a total of 24 columns out of which only 10 are useful and the other columns are dropped. All the bad rows, which consist of NULL values, zero as the value in budget and revenue columns and empty JSON values in any of the columns were removed. Once the data is cleaned 4 different tasks were performed for analysis of the above mentioned research questions.
   a) A mapping between cast ID and cast name
   b) A mapping between director ID and director name
   c) A mapping between the cast and the list of movies it has been on where only top 3 casts were considered.
   d) A mapping between the movies and the list of cast where only top 3 casts were considered in the decreasing order of their importance.
2) TMDB Dataset:
   This dataset was generated from The Movie Dataset API. This dataset provides information about different movies, namely, title, cast, crew, budget, revenue, etc. It consists of 2 files with 4803 movie information records each [6].
   - credits.csv: This contains fields like the movie_id(tmdb_id), title, cast and crew of a particular movie. The total number of record is 4813 out of which 14 rows have at least one column that is missing. There are 42 duplicate cast entries and 27 duplicate crew entries which belong to different movies.
   - movies.csv: This has fields including but not limited to movie_id(tmdb_id), the title, budget for the movie production, genre of the movie, the original language, etc. Total number of records in this file is 4806 out of which 6 have at least one column missing. 28 rows have the value for Genre missing, 1 row has release date missing, 2 rows have runtime missing, 86 rows were missing spoken languages attribute.
3) The MovieDB:
   The primary source of this dataset is www.themoviedb.org. This dataset contains information on various directors, awards for which they were nominated, the number of total awards won, etc. This dataset consists of 8 files with varying numbers of records. The files we will be using from this dataset are 900_acclaimed_directors_awards.csv and 220k_awards_by_directors.csv [7].
   - 900_acclaimed_directors_awards.csv: This presents us with details about different directors and the number of awards they received. There are a total of 6676 columns and 894 rows in this file.
   - 220k_awards_by_directors.csv: This consists of details about the award nominations that different directors received in various award ceremonies. There are a total of 6 columns and 225676 rows in this file.

   First, all the columns which were not necessary were dropped from both the tables and upon inspection no bad rows were present, no NULL values.

VI. Methodology

In this section, we discuss various approaches to answering our research questions.

A. Research Question 1

The research question for this part is "Is there a relationship between the release month and the movie's profitability and popularity? If yes, which months turn out to be most lucky for movies in terms of popularity and profit?". We perform joins on the files obtained from the "TMDB" and "The movies dataset". This is achieved using Hive [1]. In the process, we remove the duplicate, null and empty entries. Only the columns containing `movieId`, `movieName`,

releaseMonth, popularity, revenue and budget are retained. Following this, we import the data in Tableau and find the relationship using graphs.

B. Research Question 2

This part of the research will discuss "How has the time duration been affecting profits, voting and popularity of the movies over the years from 2007 to 2020?". A join is performed on the files obtained from "TMDB" and "The movies dataset". The join is performed using Hive. Duplicates, nulls and empty entries are removed as part of the process. These columns are the only ones retained: movieId, movieName, movieDuration, popularity, revenue and budget. After importing the data into Tableau, we use graphs to determine the relationship.

C. Research Question 3

In this part, we try to find the answer to the research question - "Does the popularity of a director in a particular calendar year affect the chances of them being nominated for an award at the Cannes Film Festival?". We use the data from all the sources for this one. First, a table containing movieName, movieId, releaseDate, directorId, popularity, budget and revenue is extracted from the "TMDB" and "The movies dataset". In the next step, we extract the directorId and the years in which they were nominated for Emmy awards. Both these tables are imported into Tableau and Tableau "relations" are used to plot the graphs to answer the question.

D. RoI prediction model

A schematic diagram of the model is shown in Fig. 1.

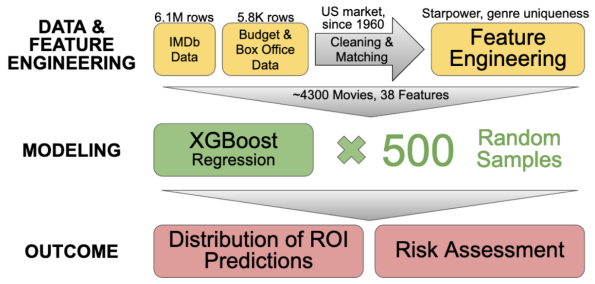

Fig. 1. Diagram outling the modelling process behind RoI prediction

1) Feature Selection and Engineering: Some of the fields used by the Return on Investment(RoI) model are budget, runtime, releaseDate and genre.
Creating useful features for the model took a substantial amount of time. A series of three metrics were used: actor starpower, director starpower and genre uniqueness. A movie's success can be attributed to the relative prominence of its actors and director. It is also possible for a movie to pique the interest of moviegoers if it is a mash-up of several genres (example: film noir + sci-fi + thriller). See Fig. 2 for an explanation of actor starpower.

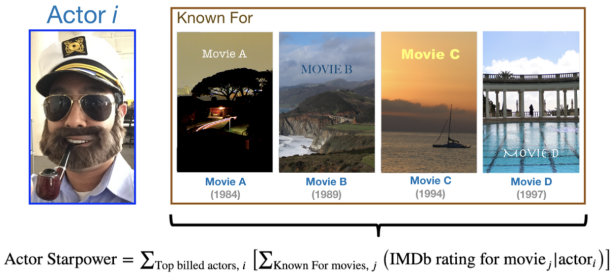

Fig. 2. Computation of actor starpower

Similar to the starpower for a single actor, the starpower for a director is also computed. Genre uniqueness measures how unique a movie's combination of genre categories is relative to all movies in the data set. You can see the formula in Fig. 3.

$$\text{Genre Uniqueness} = -\log\left(\frac{\#\text{ movies with this combo of genres}}{\text{Total \# of movies in data set}}\right)$$

Fig. 3. Computation of genre uniqueness

The "-log" here serves to create a more normally-distributed quantity while ensuring that more unique genres have a larger, positive value. Ultimately, we have 9 features, most of them categorical and one-hot encoded. The selection and engineering of features is laborious, but vital, since a successful model depends heavily on the quality and quantity of input data.

2) Building Model to Predict Movie RoI: Profitability is used as a metric of success for a film and is defined as return on investment (RoI). The RoI is simply the fraction of the budget that the movie makes back at the box office (i.e., RoI = Profit/Budget). Since extreme values of RoI are fairly common for movies (both massive successes and major flops) and the range is large, the target variable that we aim to predict is log(RoI + 1).

Using the root-mean-square error (RMSE) as the goodness metric, we selected XGBoost as the regression model since it was found to outperform random forest regression. Several parameters of the XGBoost model were tuned through 5-fold cross-validation, including the number of trees, the depth of each tree, and the learning rate.

VII. Experimental Setup

The design of the workflow is illustrated in Fig. 4.

A. Platform

NYU Peel cluster was used for storing data on Hadoop, running map-reduce jobs and hive queries.

B. Technologies and frameworks used

- HDFS
- Hive
- Tableau
- Java

C. Workflow

The data is downloaded from Kaggle and stored on Hadoop Cluster. Post this, the data went through the following phases (using map-reduce jobs):

- Profiling: This is where various characteristics of the data were discovered.
- Cleaning: Using the statistics from profiling, data cleaning was carried out where operations like null field removal, bad row removal, and column filter were carried out.
- Extraction: In this phase, the data was organized into well-formed tables for analysis. These tables were stored in Hive.
- Analysis: Various complex join queries were performed on the tables to get all the data needed for the analysis in a single place. Analysis was performed by visualizing data using Tableau.
- RoI modeling: In this phase, the columns required by the model were extracted from Hive and used for training and testing the RoI prediction model.

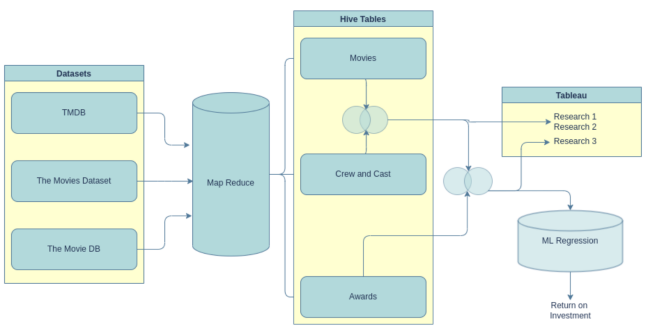

Fig. 4. Design of the workflow

VIII. Results and Observations

We have organized the results into sections. We start with the discussion of some data characteristics. This is followed by top-k analytics. We then present observations on our research questions. A case study of "Die Hard 2" movie is included in the final section along with the results of the RoI prediction model.

A. Data characteristics

Different columns of the dataset are plotted in the Fig. 5. Following are the observations:

- Most movies lie in the budget range 0 to 0.5 on a "e to the power 8" scale.
- Most movies were made in the months of January and December.
- Most movies have popularity between 0 to 100.
- Many movies have negative profit values that suggest loss-making movies.
- Most movies have revenue collection in the range 0 to 0.25 but on a "e to the power of 9" scale.
- Most movies have a runtime in the range of 75 to 150.
- Vote average has a bit more scattered distribution than other variables with most movies lying in the range of 6-7 voting average.

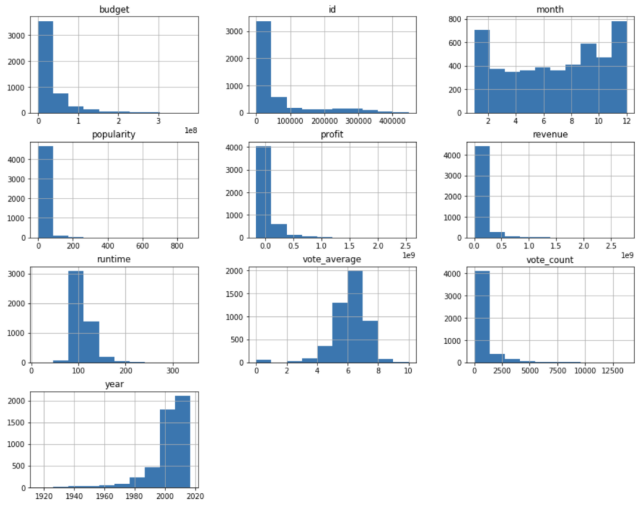

Fig. 5. Characteristics of the data

We also plot the correlation graph of the important columns. The graphs are shown in Fig. 6 with the observations listed below.

- Profit vs Popularity shows a positive but low correlation.
- Profit vs Revenue shows the highest positive correlation.
- Profit vs Runtime shows a positive but very low correlation.
- Vote Average vs Runtime also shows a positive but low correlation.
- Popularity vs runtime also has a very low but positive correlation.

B. Top-k analytics

1) Top movies by Revenue and Profit: From Fig. 7 and Fig. 8, we observe that

- The movie with the maximum profit and revenue is "Titanic" and it is first by a huge margin.
- Most of the movies which have a high revenue also have a high profit. This supports the behavior observation from the previous subsection that profit and revenue have a positive correlation.
- There are movies with high revenue but not so high profit implying a huge budget.

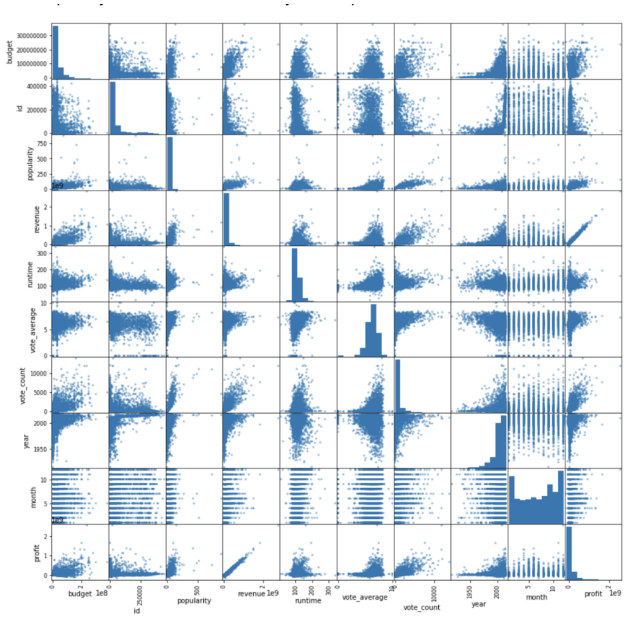

Fig. 6. Correlation chart

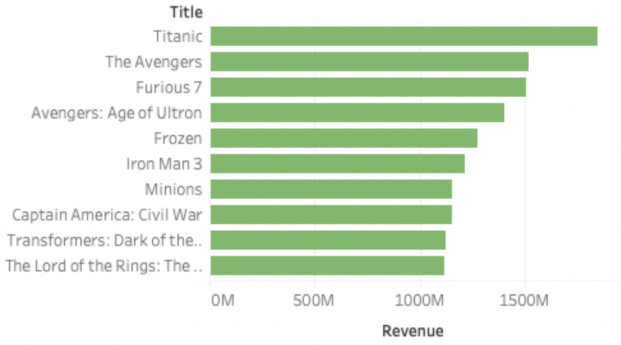

Fig. 7. Top movies by revenue

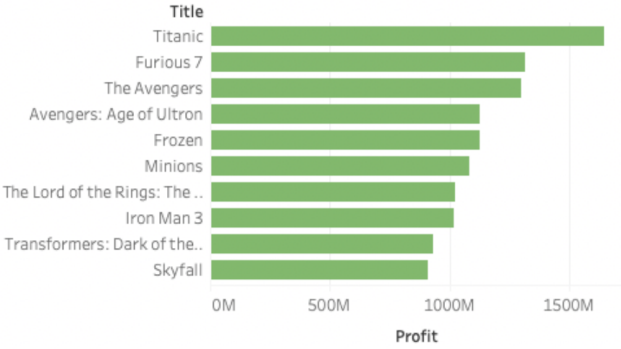

Fig. 8. Top movies by profit

2) Top movies by Popularity: Fig. 9 gives us an insight into the movies which have been most popular over the years.

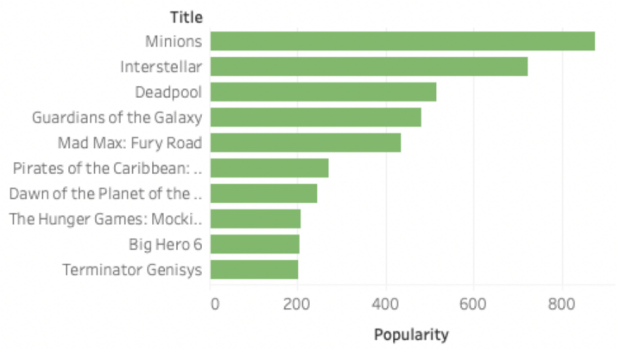

Fig. 9. Top movies by popularity

C. Research Question 1

In this section, we present insights into the movie profits and popularity with respect to the release month. Using Fig. 10 and 11, we make the following observations:

- June has the highest average profit and September (close to January) has the lowest average profit for the movies released on those days.
- May, June, July, November and December have average profits above the overall mean profit value.
- June has the highest average popularity and September (close to January) has the lowest average popularity for the movies released on those days.
- May, June, July, November and December have average popularities above the overall mean profit value.

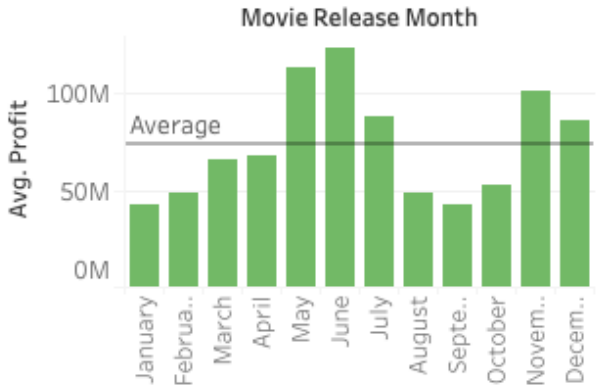

Fig. 10. Average profit for each month between years 2007-2017

D. Research Question 2

We analyze average profit and popularity as a function of the duration of a movie. From Fig. 12 and Fig. 13, we observe:

- Movies with medium duration are the most popular and have gained the most profit.
- Long-duration movies have not been profitable and their popularity has been less as well.
- Short movies haven't been much popular and have resulted in a below-average profit.

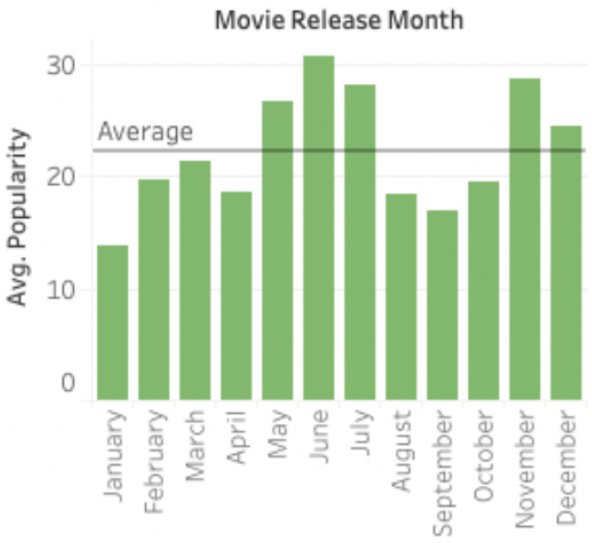

Fig. 11. Average profit for each month between years 2007-2017

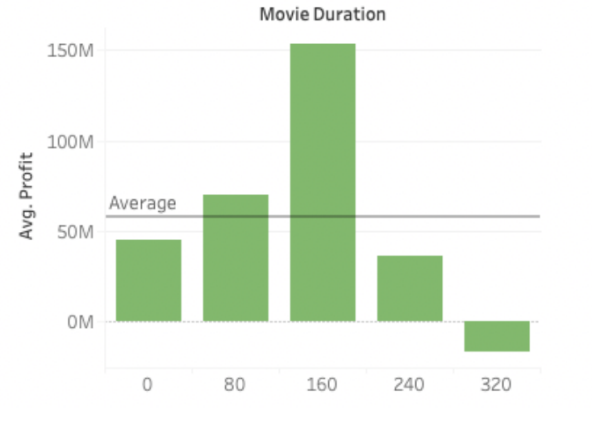

Fig. 12. Average profit for movie duration buckets

We also analyze the variation of the popularity and profit as a function of movie duration to observe the trend over years using Fig. 14 and Fig. 15.

- Maximum popularity on an average is attained by medium films in 2014.
- Generally, medium films have gained more popularity than other category films and have shown increasing average popularity trends over the years.
- Long films have been the lowest in terms of popularity over the years.
- Over all the years, medium films have gained average popularity more than the overall popularity mean.
- Short films have shown increasing average popularity trends over the years with 2015 and 2016 being the years crossing the overall popularity mean.
- Long films data is not present for 2011 - 2015 and 2017.

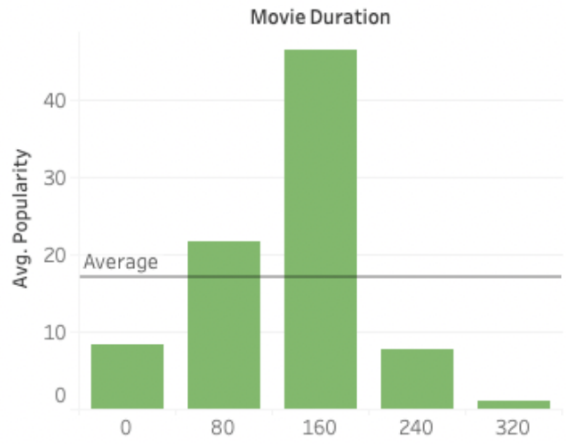

Fig. 13. Average popularity for movie duration buckets

- Maximum popularity on an average is attained by long films in 2009 followed by medium films in 2017.
- Short films data is not present for 2017.
- Short films have always been below overall mean vote rating on an average for all the years.

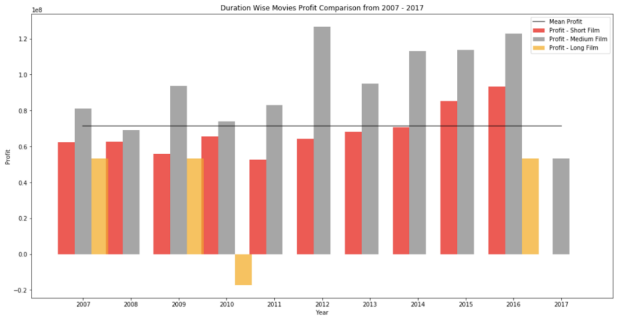

Fig. 14. Time series chart of movie duration v/s profit

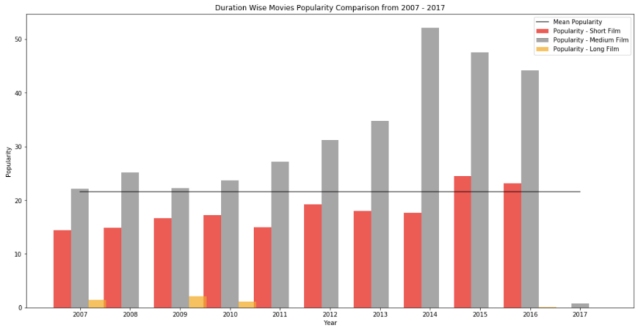

Fig. 15. Time series chart of movie duration v/s popularity

E. Research Question 3

In order to find a relation between the popularity of a director and them being nominated for an Emmy award in a particular calendar year, we plot the graph shown in Fig. 16. The red dots show the popularity of the movies nominated for Emmy and the green ones is the set of popularities of all the movies. The figure contains data of all the directors. No correlation can be seen between the popularity and the Emmy nomination.

Fig. 16. Director popularity(a.k.a. starpower) as a function of year

F. RoI model

The result of a single XGBoost model trained on 80% of the data and tested on the unseen held-out 20%.

The scatterplot in Fig. 17 is proof that it is hard to predict the success of movies! One model's prediction of RoI output would not be very accurate. Using random subsamples of the training data, one can create a distribution of RoI predictions that serves as a proxy for the amount of risk involved in funding a movie. Based on the complete training set of N samples, we generated 500 subsamples each of size N/2 that were randomly selected from the full set of N. 500 and N/2 are somewhat arbitrary values that were chosen to obtain a smooth distribution of RoI values and to maintain a sufficient training set for each model. From these 500 random subsamples, we trained 500 models and built a distribution of RoI values from which we can derive summary statistics such as the median and 95% confidence interval.

G. "Die Hard 2" analysis

In Fig. 18, the blue histogram represents the distribution of the 500 RoI predictions derived from the pre-trained XGBoost models. Those areas shaded in green represent a positive profit regime (RoI $>$ 0%), and those in gray represent losses. "Die Hard 2" has almost all RoI projections falling into the "profit" regime, which means that the project is "SAFE" and it can be funded. One can select what percentage of RoI predictions can fall in the "loss" regime while still considering the movie a "SAFE" investment through the risk tolerance setting. Movies not classified as "SAFE" investments are labeled as "RISKY"

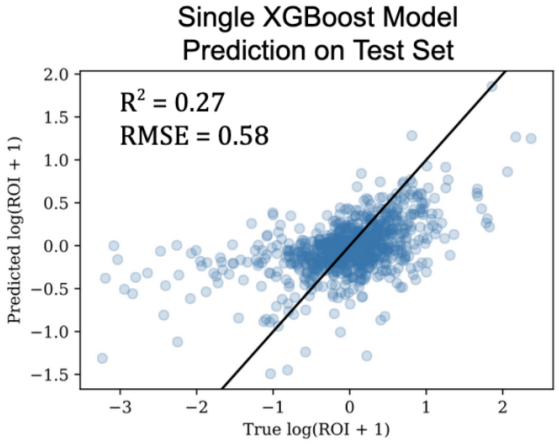

Fig. 17.

investments, and an example distribution of each is shown in Fig. 19.

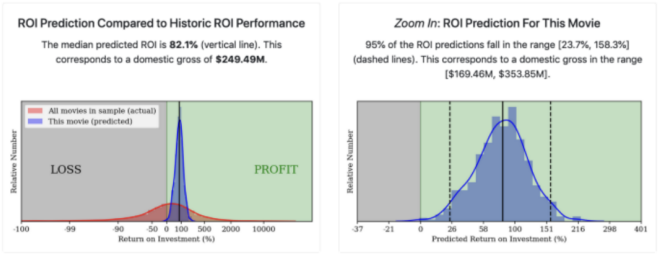

Fig. 18.

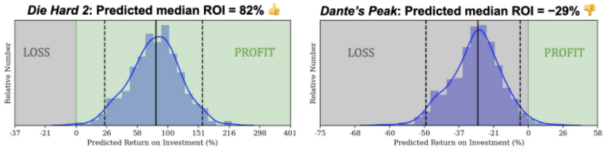

Fig. 19.

IX. Goodness

There a number of factors that assess that the analytic are good. These are listed below.

- We use multiple movie and awards datasets.
- Data segmentation is performed which is based on the release year of the movie. This is done because parameters like popularity, which is based on the number of tweets, are not a true reflection of a movie's success before twitter came into use as a major microblogging and social networking service.
- The year of release vary from Golden Age Era(1950s) to 2000s because of which we convert the movie budget and revenue to 2019 USD using the CPI data.

X. Coding Challenges

We faced a few challenges while working on this project. The solutions to the challenges are also mentioned along with the challenges.

- The first challenge we faced was to map the cast of a movie to a single number which is then used in RoI prediction model because there are a lot of deciding factors for a movie's generated revenue. We solve this by taking the average votes of all the movies the case has been a part of. This cast rating can be determined by using the formula in Fig. 20.

$$\sum_{\text{Top billed actors, } i} \left[\sum_{\text{Known For movies, } j} \left(\text{IMDb rating for movie}_j | \text{actor}_i \right) \right]$$

Fig. 20.

- The second challenge was to find duplicate movie titles. For example, the original Star wars movie is listed as both "Star Wars" and "Star Wars: Episode IV - A New Hope". Bloom filter is a space-efficient probabilistic data structure that is used to test whether an element is a member of a set. It doesn't generate a false negative result. We use bloom filter as solution for this challenge and search for common words in the filter.
- Mapping all the genres to their respective uniqueness number based on their importance which is used in RoI prediction model was our third challenge. We use the formula mentioned in Fig. 3 to obtain the genre uniqueness. The "-log" in the formula serves to create a more normally-distributed quantity while ensuring that more unique genres have a larger, positive value as mentioned before.

XI. Obstacles

We faced the following obstacles while finding the correlations between attributes.

- Designing the logical tables to find a relationship between a director's popularity in a given year and whether they were nominated for Emmy awards for that award year. This was solved using `relations` in `Tableau` where we can combine data in a flexible way for the analysis.
- Obtaining the popularity of a director in a particular year and plotting the popularity for the year in which they were nominated for an Emmy award. We designed hive queries carefully to overcome this obstacle.

XII. Conclusion

From the analytic we can deduce that Movies that were released around holiday time, months July and December, make the most profit and profits were lowest during the months January and September. Medium duration movies attain maximum profit on an average as compared to short duration or long duration movies where long duration movies were on an average at a loss. Medium duration movies have been slightly more popular and are gradually gaining more popularity along with short duration movies. We can also concur that there is not any correlation between a director's popularity and them getting nominated for an Emmy Award in a particular calendar year. The results from the RoI model also indicate that the success and the revenue of a movie cannot be predicted.

XIII. Acknowledgement

We would like to express our special appreciation to Prof. Tang as well as the HPC team for providing the guidance and the resources. We are grateful to the Tableau software as well which made the analysis process smooth and easy for us.
Additionally, a big thanks to the open source technologies by Apache without which this work would have never been possible.

References

[1] "Predicting Gross Movie Revenue", https://arxiv.org/pdf/1804.03565.pdf
[2] "Big Data and Hollywood: A Love Story", https://www.theatlantic.com/sponsored/ibm-transformation-of-business/big-data-and-hollywood-a-love-story/277/
[3] "How Data Science Is Used Within the Film Industry", https://www.kdnuggets.com/2019/07/data-science-film-industry.html
[4] J.S. Simonoff and I. R, Sparrow, "Predicting movie grosses: Winners and losers, blockbusters and sleepers", Stern School of Business, New York University, 1999, pp. 99-8. http://hdl.handle.net/2451/14752
[5] "The Movies Dataset", https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset
[6] "TMDB 5000 Movie Dataset", https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata
[7] "The MovieDB", https://www.kaggle.com/datasets/stephanerappeneau/350-000-movies-from-themoviedborg
[8] "Apache Hive", https://hive.apache.org/