

Box Office Revenue Prediction

Project Proposal Paper

Nhan Nguyen

Computer Science

University of Colorado Boulder

Colorado United States

Nhan.N.Nguyen@colorado.edu

Melissa Pollich

Computer Science

University of Colorado Boulder

Colorado United States

Melissa.Pollich@colorado.edu

Benjamin Breyfogle

Computer Science

University of Colorado Boulder

Colorado United States

Benjamin.Breyfogle@colorado.edu

PROBLEM STATEMENT/MOTIVATION

As of 2018, it has been reported that the global box office is worth \$41.7 billion. However, when the home entertainment revenue is included, the global film industry is worth over three times that at \$136 billion. Despite the massive profits, movie industry is one of the riskiest markets for investors due to its uncertainty and unpredictability. Once a movie fails to meet expectation, it can potentially place a stress on the financial status of the movie studio, and lead to the withdrawal of funds from the investors. Therefore, through analyzing the performance of movies on the market, we could find the correlation between attributes of a movie and predict the revenue on release.

CCS CONCEPTS

• **Computing methodologies** → Retrieval model and ranking

KEYWORDS

Revenue, profitability, variables, data

ACM Reference format:

N. Nguyen, M. Pollich and B. Breyfogle. 2018. Box Office Revenue Prediction : Project Proposal Paper. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18)*. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/1234567890>

*Article Title Footnote needs to be captured as Title Note

†Author Footnote to be captured as Author Note

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WOODSTOCK'18, June, 2018, El Paso, Texas USA

LITERATURE SURVEY

Galvão, M. and Henriques, R. (2018). Forecasting Movie Box Office Profitability. *Journal of Information Systems Engineering & Management*, 3(3), 22. <https://doi.org/10.20897/jisem/2658>

The revenue of a movie can be predicted by constructing a predictive model based on regression, decision trees or neural networks. The paper proposed 3 different variables to use: an interval variable for the value of profitability of individual movies, a categorical variable with multiple classes of profitability values, and a binary variable that shows whether the movie is profitable or not. Using these three variables can increase the accuracy of the prediction and help answer how the market behaves, what appeal to the audiences and the risk associated with movies production.

PROPOSED WORK

Overall, our proposed methodology has three components: Data Collection, Data Cleaning, Data Analysis and Prediction. Cleaning the dataset and discarding the irrelevant data from the datasets as well as thorough analysis of the data within these sets. We will compare the attributes that are provided across all of the datasets and build a brand a database that encompasses only the attributes that we care about.

1. Data collection: We plan to utilize three different datasets that all contain different attributes for all the

movies that were produced within the last decade or so. Within these selected datasets, some common attributes included: ratings (IMDB), title, actors, actresses, studio, and release date. We will explore in more detail which attributes we want to focus on in the Evaluation Methods section of this paper. Lastly, we chose not to collect data from the IMDB dataset due to multiple reviews expressing the difficulty with the formatting of the data.

2. Data Cleaning: In order to clean the data from all of the datasets, we will construct a relational database to make finding duplicates and unnecessary attributes easier. Moreover, with this database we will be able to write various SQL queries in order to narrow down relevant information.

3. Data Analysis and Prediction: After the data is cleaned and organized efficiently, we will analyze it and look for distinguishing patterns. As proven to be effective in the surveyed literature, our prediction model will use a classification system in order to separate the movies based off of the attributes that contribute the most to a higher box office revenue.

DATA SET

Boxofficemojo Alltime Domestic Data

The data contains 16223 unique values of the lifetime gross, ranking, title, studio and production year of Hollywood movies. They were scrapped from BoxofficeMojo's listing and based on domestic gross.

<https://www.kaggle.com/eliasdabbas/boxofficemojo-alltime-domestic-data/version/3>

The Movies Dataset

This dataset contains files containing metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of 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. This dataset also has

files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

<https://www.kaggle.com/rounakbanik/the-movies-dataset/downloads/the-movies-dataset.zip/7>

MovieLens Dataset

This dataset contains six files: genome-scores.csv, genome-tags.csv, links.csv, movies.csv, ratings.csv and tags.csv and describes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service. It contains 20000263 ratings and 465564 tag applications across 27278 movies. These datasets were created by 138493 users between January 09, 1995 and March 31, 2015 and was generated on October 17, 2016. Users were selected at random for inclusion. All selected users had rated at least 20 movies. No demographic information is included. Each user is represented by an id, and no other information is provided.

<https://grouplens.org/datasets/movielens/20m/>

EVALUATION METHOD

We are looking for a high percent of support and confidence between revenue and studios, budgets, genre...

We will prepare a training set and a testing data set for our analyses. This will allow us to determine our models' accuracies.

We will be able to evaluate the relative contributions of different factors affecting movie revenue at the box office.

Most of the attributes that we will use in the project are adopted from previous studies that have found especially useful and relevant. Moreover, we will consider not only the most relevant studies with a higher success rate (sequel, actor, director, budget, nominations, genre), but also other attributes that are not as commonly considered (reviews, duration,

number of screens). Some of the attributes that we have found to be used across all studies include:

Budget

Previous studies have shown that the budget of a movie is strongly correlated with the predictability of its popularity and reception. However, while some argue that big budgets represent better quality and positive reviews from both the audience and movie critics, the studies have shown that a big budget does not ensure a high revenue. Despite this finding, this attribute could still prove valuable in our analysis and design of the prediction model.

Genre

While many authors and critics state through their studies that action movies and thrillers are the most significant and popular among audiences, the evaluation of each genre and its correlation to the film's overall success will ultimately help weed out variables that are not relevant.

Premier Date

The movies premiere date is a variable that influences box office revenue. This may be due to the fact that movie attendance, overall, tends to increase significantly on holidays or festive seasons.

TOOLS

Python – Pandas, Dataframe, Numpy, Sklearn

Github

MILESTONES

Running codes to be implemented by next week

Graphs of the data set by next weeks

Data integration by following weeks

TBD

ACKNOWLEDGMENTS

REFERENCES

- [1] Galvão, M. and Henriques, R. (2018). Forecasting Movie Box Office Profitability. *Journal of Information Systems Engineering & Management*, 3(3), 22. <https://doi.org/10.20897/jisem/2658>
- [2] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4, Article 19 (December 2015), 19 pages. <http://dx.doi.org/10.1145/2827872>