CSCI 4502 Project Part 2

Project Proposal Paper

Ben Breyfogle
Computer Science
University of Colorado Boulder
Boulder, Colorado USA
benjamin.breyfogle@colorado.edu

Nhan Nguyen
Computer Science
University of Colorado Boulder
Boulder, Colorado USA
Nhan.N.Nguyen@colorado.edu

Melissa Pollich
Computer Science
University of Colorado Boulder
Boulder, Colorado USA
Melissa.Pollich@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.

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

1. Data cleaning: Remove attributes that will not be used in mining
2. Data preprocessing: Normalize data and join across datasets for one complete set with which to work
3. Data integration: possibly include other data sources and join into one

DATASET

Boxofficemojo Alltime Domestic Data

The data contains 16223 unique values of the lifetime gross, ranking, title, studio and production year of Hollywood movies through 2018. They were scraped 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.

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/

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. DOI=http://dx.doi.org/10.1145/2827872

EVALUATION METHODS

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

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

TOOLS

1. Github
2. Python with dataframe, pandas, numpy, sklearn

MILESTONES

Running codes by next week

Graphs of the data set by next weeks

Data integration

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