Predicting Box Office Revenue

IIT2013196,IIT2013128,IIT2013172

6 September 2016

1 Introduction

Film industry is a big business in United States.It is one of the biggest player in the entertainment industry.Predicting the Gross revenue movie would be able to generate could be great help to the producers.Till now there is no computational model that can effectively predict the Gross revenue movie will be able to collect.This depends on a lot of factors like release date,number of available theatre screens,budget of the film, etc.

2 Literature Survey

A Lot of factors that affect the revenue prediction have been studied by different researchers.

Do Critical Reviews Really Matter? As mentioned in [7] even though some reviewers may have different tastes than the people reading their reviews, consumers still read the reviews before going to watch the movie to know whether it is worth their money or not. This tells us that critic ratings do matter in the revenue prediction model.

Forswell in [9] collected public information available on IMDB site for movies after January 1st, 1990 and only included those having total box office revenue greater than 100,000 dollars totaling 2500 movies.He then used linear regression model using features first week end revenue,budget and number of available theatre screens.It did not give very good result and when he tried splitting the dataset, number of training samples decreased and performance was reduced.

Robert in [11] divided the set of features in simple , complex and sentiment where simple is numeric only, complex is numeric and text based and sentiment includes all.He thought that this problem can be modeled both as regression and classification. He used logistic regression for classification by making classes by dividing the range of min and max value of gross revenue into buckets.After doing more analysis he found that text and sentiment features were insufficient to predict the revenue.

Some researchers like in [3] tried to predict the revenue before release or after 1st week of revenue. They analysed different features like production budget of the film,revenue of the 1st weekend , sequel movie or not,star power,MPAA(Motion

Picture Association of America film rating system) rating etc. He analysed the change in gross revenue with respect to rating of the movie.He tried to give weights to these categorical variables and used linear regression to do the prediction.

Vitelli in [14] tried to create a set of features and did extract values from graphical properties of the actor-actor, actor-movie, and movie-movie relationships.

Some tried to predict the revenue before the release and some after 1st week of release.Budget of the film seems to different effects on different genre movies. Mostly big budget films which are Action,Animation, Adventure tend to attract a large amount of audience whereas on the other hand movies which even after spending a lot of money like which are biography,drama,etc tend to earn less.This fact has been supported in [5].

Genre and revenue analysis was done. In Anast [1] who tried to showcase some relationship between genre and revenue. Prag and Casavant [2] showed a negative relationship for drama and revenue.

Pre release revenue Prediction although looks attractive but it is a very difficult job.Extracting information from a variety of sources like blogs,youtube trailers comments,estimating the people's response before the movie is released.Use the number of viewers and users who edit the page in wikipedia was used by Marton in [8] to find or estimate the popularity parameter in his model to find the pre release revenue. To find the estimate after 1st week gives us more predictability as it is almost twenty to twenty five percent of the gross domestic revenue. It gives us insight that we will get more accurate or better results if we estimate it after 1st week of release or Post Release.

Dursun Delen in [10] used Neural Network with features like MPAA rating,genre,star value,sequel,special value but for prediction of pre-release revenue. It was used because neural networks can handle a mix of continuous and discrete values pretty well.

Some people also tried to predict the pre release box office revenue using movie scripts as in [6] and also tried using the tweets data as in [15], blog data to anaylse the behavior of the people but these methods did not give satisfactory results.Thorsten in [12] also anaylses the negative and postive effect of tweets on the behavior of people which in turn affects the box office revenue.

Till now everyone was concentrating on Domestic revenue only but researchers as in [4] tried to estimate the Foreign Revenue the movie will be able to collect. This depends on a lot of factors like domestic success of the movie , language adaptability, cultural differences,MPAA rating differences between domestic and foreign country.

Till now all the researchers were considering all the different factors that affect the movie success.But Thorsten in [13] did a study on inter dependence between these features whether success or effect of one feature can affect the effect of other on revenue.For example,does advertising influence the box office revenue directly by creating a media presence, or indirectly through impacting consumers quality perceptions of the movie?

Overall all the researchers tried to estimate the Gross Box office revenue

movie will be able to collect.Some considered it to be regression problem and few as classification problem.None of them tried to analyse it genre specific. As if saying that all these features like MPAA rating , production budget of the film, marketing budget of the film,critic rating in the first week are not all important for all genre. Biography or Drama movie may be good spending a lot of money but would not be able to generate a lot of revenue.Likewise, only action and animation big Budget films attract a lot of audience not true for other. Some superhero movies like Super Man and others saying Big Action Movies people don't care about critic ratings they just see it. We should not consider Critic Rating in estimating Action movie revenue. And , budget of the film should not a parameter in Drama or Biography films. Talking about MPAA ratings, movies that are PG-13 , R and action,adventure,horror affect the revenue. They should be considered as a parameter. Whereas movies made for children like G rated , they when released in US mostly collect almost equal amount of revenue.Here we are focusing on post release domestic revenue prediction only.

3 Table

Name	Information
Algorithm and tools	Regression, Classification, Neural Network.
Set of features	movie rating, critic views, theatre screens,budget,release date.
DataSet	Rotten Tomato, Wikipedia, IMDB.
Genre	Sci-Fi, history, thriller,horror, comedy, cartoon, action, documentary
Types of Revenue	Domestic Gross, Foreign Gross
Types of Prediction	Pre-release and Post-Release(After 1 week)

4 Plan of work

We will estimate the domestic gross revenue movie will be able to collect.

4.1 Steps

1. Data Collection
2. Data Cleaning(Preprocessing)
3. Data Transformation (Normalization)
4. Estimation Using Regression and Classification Models

4.2 Data Collection

We don't have a dataset. We will crawl and collect data about different movies from different websites like IMDB,MetaCritic,Rotten Tomoatoes, BoxOffice-Mojo. All the data collected is then merged into one to make a database

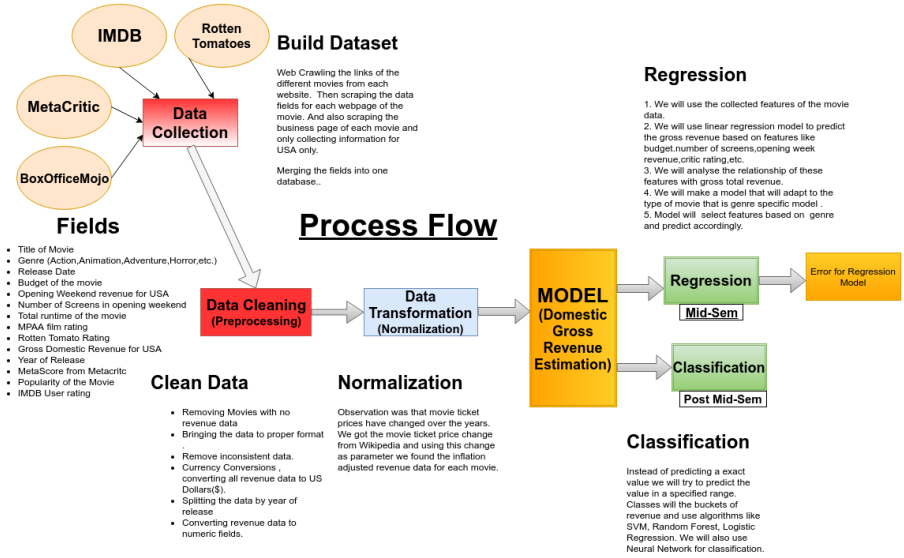

Figure 1: Process Flow

containing all the information about the movies. We also scraped the business page for each movie and collected revenue information for only USA. We are focusing on USA region only.And the gross revenue is also domestic total gross revenue for USA region only.

We collected the data from year of realease from 2000 to 2015 only.

Fields:

1. Title of Movie
2. Genre (Action,Animation,Adventure,Horror,etc.)
3. Release Date
4. Budget of the movie
5. Opening Weekend revenue for USA
6. Number of Screens in opening weekend
7. Total runtime of the movie
8. MPAA film rating
9. Rotten Tomato Rating
10. Gross Domestic Revenue for USA

11. Year of Release
12. MetaScore from Metacritic
13. Popularity of the Movie
14. IMDB User rating

4.3 Data Cleaning

1. Removing Movies with no revenue data
2. Bringing the data to proper format .
3. Remove inconsistent data.
4. Currency Conversions , converting all revenue data to US Dollars
5. Splitting the data by year of release
6. Converting revenue data to numeric fields.

4.4 Data Transformation

We observed that movie ticket prices have changed over the years.Due to inflation price of tickets have changed so much over the years. We got the movie ticket price change from Wikipedia.We used this change in ticket price and considered this as inflation parameter and calculated the inflation adjusted revenue data for each movie.

Ticket prices changed so rapidly with price of 5.39 USD in year 2000 to about 8.66 USD in 2016.

4.5 Estimation Using Regression Models

We will use regression to find the approximate value of the gross domestic revenue for the movie.We will use different combination of features and test what effect the gross revenue most.

1. We will use the different features of the movie data.
2. We will use linear regression model to predict the gross revenue based on features like budget.number of screens,opening week revenue,critic rating,etc.
3. We will analyse the relationship of these features with gross total revenue.
4. We will make a model that will adapt to the type of movie that is genre specific model .
5. Model will select features based on genre and predict accordingly.

6. Making a model that will change its set of input parameters required based on type of movie .

7. Calculate error for each type of movie for each regression algorithm.

4.6 Estimation Using Classification Models

Instead of predicting a exact value we will try to predict the value in a specified range. Classes will the buckets of revenue and use algorithms like SVM, Random Forest, Logistic Regression. We will also use Neural Network for classification.

This part will be done post mid semester.

References

[1] *Anast, P. 1967. 'Differential movie appeals as correlates of attendance', Journalism Quarterly.*

[2] *Prag, J.J. Casavant, J. 1994. 'An empirical study of the determinants of revenues and marketing expenditure in the motion picture industry', Journal of Cultural Economics,.*

[3] *Simonoff, J. S. and Sparrow, I. R. Predicting movie grosses: Winners and losers, blockbusters and sleepers. In Chance, 2000.*

[4] Ryan Compton Brian de Silva. *Prediction of Foreign Box Office Revenues Based on Wikipedia Page Activity.* 2014.

[5] Luis Cabral and Gabriel Natividad. *Box-Office Demand: The Importance of Being.* 2013.

[6] Z. John Zhang Jehoshua Eliashberg, Sam K. Hui. *Assessing Box Office Performance Using Movie Scripts.* 2014.

[7] Alec Kennedy. *Predicting Success: Do Critical Reviews Really Matter?* 2010.

[8] Janos Kertesz Márton Mestyan, Taha Yasseri. *Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data.* 2013.

[9] Anahita Sidhwa Nikhil Apte, Mats Forssell. *Predicting Movie Revenue?* 2011.

[10] Dursun Delen Ramesh Sharda. *Predicting box-office success of motion pictures with neural networks.* 2006.

[11] David Cummings Steven Yoo, Robert Kanter. *Predicting Movie Revenue from IMDb Data.* 2011.

[12] Fabian Feldhaus Thorsten Hennig Thurau, Caroline Wiertz. *Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies*. 2014.

[13] Gianfranco Walsh Thorsten Hennig-Thurau, Mark B. Houston. *Determinants of Motion Picture Box Office and Profitability: An Interrelationship Approach*. 2008.

[14] Matt Vitelli. *Predicting Box Office Revenue for Movies*. 2012.

[15] Ross Maciejewski Yafeng Lu, Feng Wang. *Business Intelligence from Social Media:A Study from the VAST Box Office Challenge*. 2014.