ORIE 4741 Midterm Project Report

By Song Tang (st883), Wenchang Yang(wy286), Jia Rao (jr2254)

Project Name

Box Office Prrdiction Based on Movie Reviews and Basic Profiles

Introduction

In this project, we are going to predict a movies total gross based on its review-related variables collected within the first week after the movies wide release date and other non-review-related variables. We have finished the data collection and almost done with the data cleaning part.

Dataset

We now have our data scraped from Omdb API, Metacritics and Box Office Mojo websites. We got the first-week reviews from Metacritics, total gross and opening gross from Box Office Mojo and non-review-related characteristics of movies (i.e. year, studio, runtime, rate, production budget..) from Omdb API.

We found that some movies have their opening date earlier than the wide-release date; with the wide release date meaning that the movie is released in more than 600 theaters, these movies are released in significantly less theaters on their initial/limited release date (they were usually first released on film festivals and selected art theaters). Considering the fact that we will use the first-week data as important features, for those movies with limited releases before wide releases, we modified our original dataset to use the first-week reviews as well as first-week gross after the wide release date to have a more accurate estimation (because their box office gross within first week after wide-release date can better reflect their market value).

Also, we are getting rid of those movies with total theater numbers less than 600 after their wide release dates. Besides, a small portion of re-released movies (i.e. *Titanic 3D*) will be deleted from dataset as we believe their major box office came from their initial release year.

Variables

Current review-related variables within the first week after the movies release date:

Metascore:
The metascore from Metacritic in the period of the first week after the movies release date for now.

This variable is an integer value ranging from 0-100. Each movie on metacritic has a unique metascore computed by scores given by metacritics approved reviewers (critics) based on metacritics review score system. The higher value of this variable indicates higher quality perceived by the reviewer.

In the future, we will expand this category into a set of 5 variables:

Metascore:
The variable we currently have.

Meta-userscore:
The user-review score from metacritic in the period of the first week after the movies release date. This variable is 1-digit decimal value ranging from 0.0-10.0. Each movie on metacritic has a unique user review score computed by scores given by public users based on metacritics review score system. The higher value of this variable indicates higher quality perceived by the user.

IMDb-userratings:
The user ratings score from IMDb in the period of the first week after the movies release date. This variable is a 1-digit decimal value ranging from 0.0-10.0. Each movie on IMDb has a unique user rating score computed by scores given by public reviewers based on IMDbs user rating score system. The higher value of this variable indicates higher quality perceived by the user.

RT-allcritic:
The all critics score from Rotten Tomatoes in the period of the first week after the movies release date. This variable is an integer percentage value ranging from 0

RT-topcritic:

The top critics score from Rotten Tomatoes in the period of the first week after the movies release date. This variable is an integer percentage value ranging from 0

Current Non-Review-related variables:

Year: Our dataset contains all the 1743 movies with the Total-theater number larger than 600 theaters from 2006 to 2016 based on our wide-release assumption.

Opening-gross: The box office gross within the first week after the wide release date.

Num-openingtheaters: The number of opening theaters within the first week after the wide release date.

Runtime: An integer variable indicating the duration of the movie.

Rate: The rate of the movie assigned by MPAA (Motion Picture Association of America). We will make it a dummy variable.

Production-Studio: There are six major production studios, which are called "Big Six" in Hollywood. We set this variable as a binary dummy variable: 1 if a movie is produced by one of the "Big Six", and 0 if not.

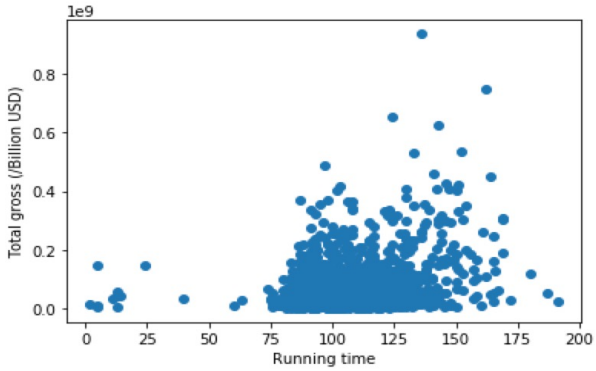
Is-usa: A binary variable showing whether or not the movie is produced by American film companies.

In the future, we will have more non-review-related variables containing:Director, Actor, Production-budget

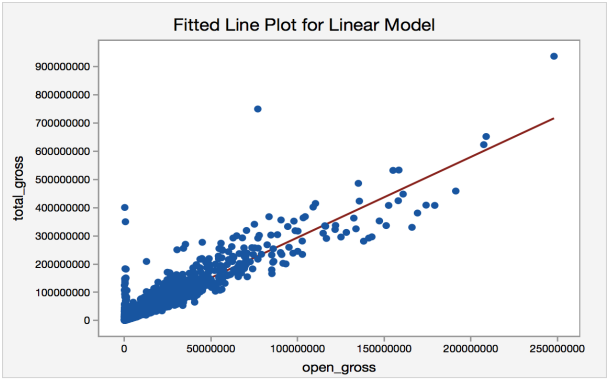
Regression

Based on the data we have, we ran some preliminary analysis. Specifically, we now have some graphical data summaries displaying linear relationships between different variables. Also, we have tried to apply the linear regression model (least squares) to those x variables shown in the graph; for now, the opening gross within the first week after movies wide release dates is the most significant one.

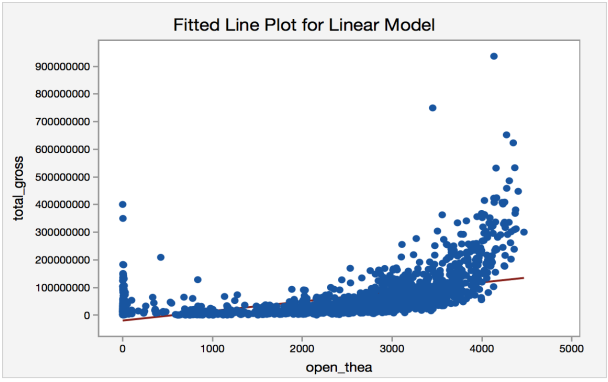
(a) The total box office grosses of movies produced by the "Big Six" studios and movies produced by other studios



(b) Total box office grosses of movies with respect to their durations



(a) Prediction of total box office grosses based on Num-openingtheaters



(b) Prediction of total box office grosses based on Opening-gross

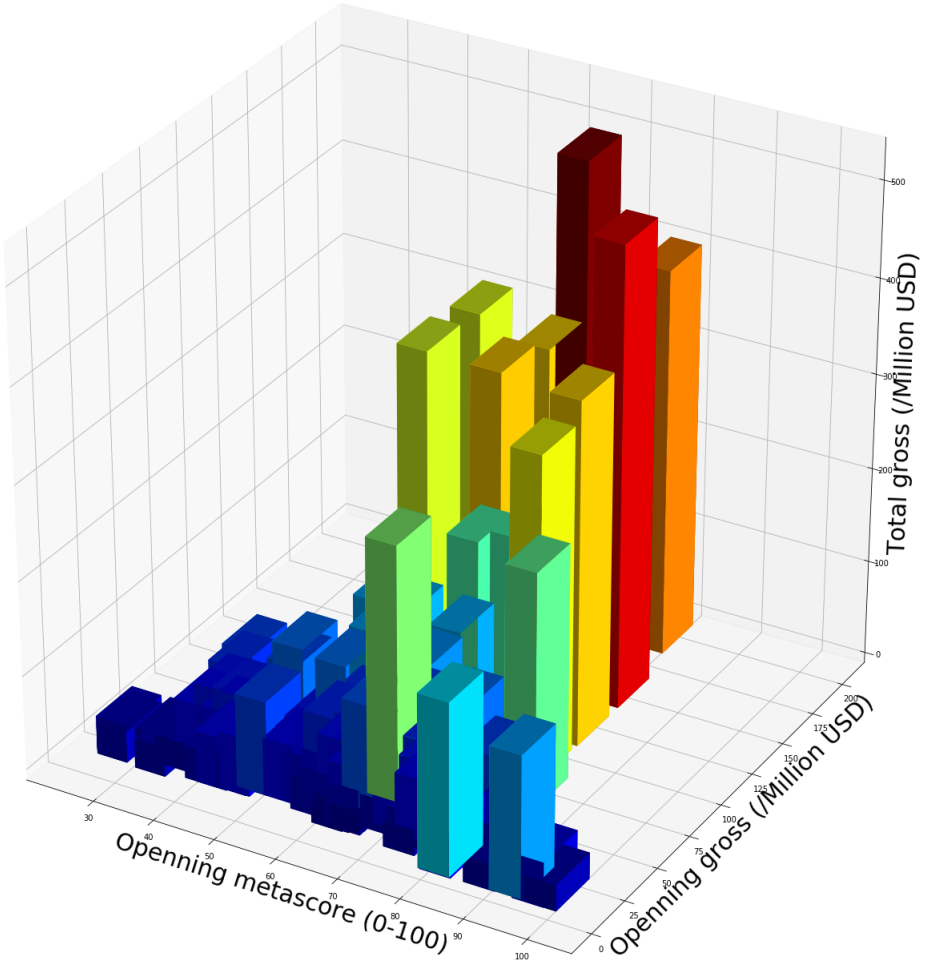


Figure 3: Total Gross with respect to Opening Metascore and Opening Gross

Future plan

The most important task is that regarding our dataset, we need to get more data to have more features to explore; Review is one of the most important features we want to use, and besides the reviews by critics from Metacritic website we have now, we need user reviews from Metacritic as well as Imdb and rotten tomato to have a more comprehensive view.

As we are collecting data from different websites to make up our own dataset, we have to make sure that our dataset is accurate with no missing values. When we were trying to apply some models to our current dataset, we found that for some specific variables there still exist a few inaccurate or NaN values. We will finish our data cleaning after we have all the data integrated.

For the feature engineering, we have selected and got most of the essential features we would like to have, and well add more to our input space to see whether we should keep them as significant ones or discard them to prevent overfitting.

At last, well apply more complicated models including the regularizers and do cross-validation to choose the best one among them.