ORIE 4741 Project Plan

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

Project Name

Box Office Prrdiction Based on Movie Reviews and Basic Profiles

Project Idea

Today, people tend to read movie reviews on the websites like Rotten Tomatoes to decide whether one movie is worth going to watch or not. Most likely, the online reviews, especially those coming out during the first week, would influence the box office performances a lot. As the reviews from the first week are the earliest and freshest, we can expect that each of them has effect on the decisions of potential audiences.

For the project, we would like to take the basic profile of one movie (i.e. category, studio brand, budget, celebrity effect), along with its first weeks online reviews (i.e. from professional critics and audiences), to construct a model to predict the box office of it.

Questions to investigate

The main purpose of our project is to predict the box offices based on the basic profile of movies and the first week's reviews. We can divide the purpose into three questions:

1) How can we use the basic profile of movies and the first week's reviews to predict their box offices?

2) How does the first weeks reviews influence the total box office?

3) Is the influence the same ten years ago? Can we apply the model derived from recent years data to the data ten years ago?

As we have the response variable (box office) with numeric value, and there exist linear relationships between the explanatory variables and the response variable (e.g. we can image that lots of good reviews during the first week will probably lead to high box office), we assume that linear models would fit the problem well. Feature engineering will also be applied.

Datasets to use

Based on the research questions, we have found three main datasets to use:

1) Movie profile dataset (Kaggle):
https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

2) Movie review dataset (Rotten Tomatoes API):
https://developer.fandango.com/Rotten_Tomatoes

3) Movie box office dataset
http://www.the-numbers.com

There are also other datasets we may explore:

1) IMDb + Rotten Tomatoes dataset for movies from 1997 to 2009:
http://wiki.urbanhogfarm.com/index.php/IMDb_%2B_Rotten_Tomatoes

2) the open movie database:
http://www.omdbapi.com