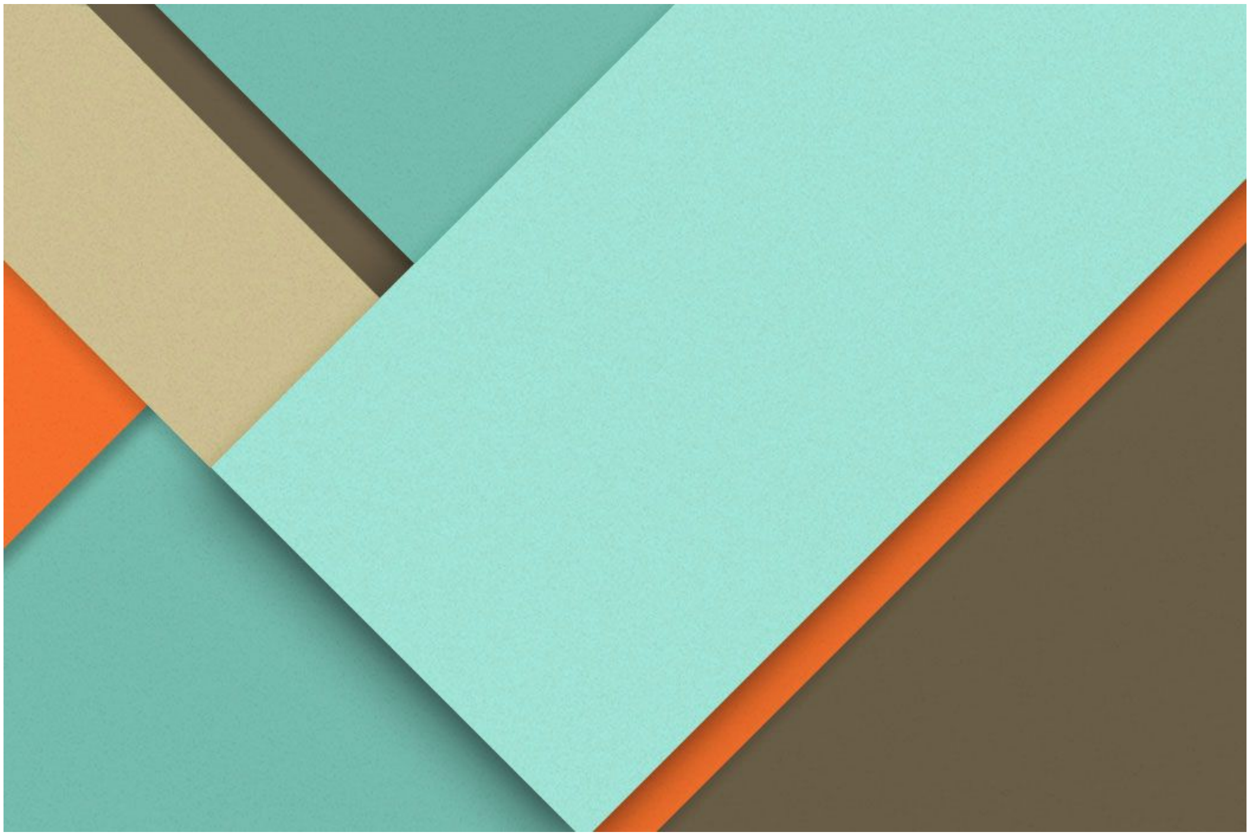


Movie Prediction

09.23.2016

Mayank Thirani (20352032)

Ryan Zhu (20356169)

University of San Francisco
2130 Fulton Street
San Francisco, CA

Contents:

Overview

Given that thousands of movies were produced each year, is there a better way for us to tell the greatness of movie without relying on critics or our own instincts?

Many people rely on critics to gauge the quality of a film, while others use their instincts. But it takes the time to obtain a reasonable amount of critics review after a movie is released. And human instinct sometimes is unreliable.

Goals

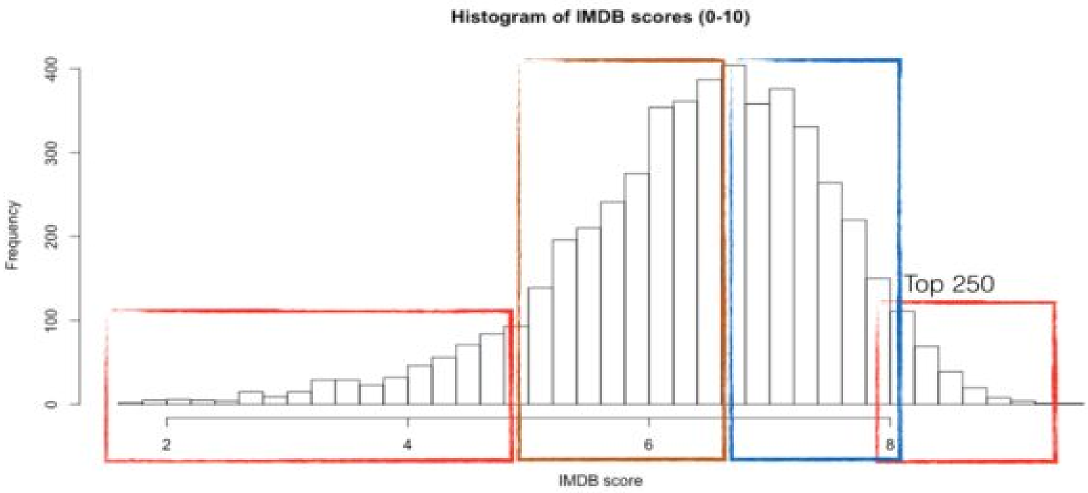
1. Predict the greatness of a movie before its released date

Problem Definition

How can we tell the greatness of a movie before it is released in cinema?

We are going to predict the movie's greatness by the movie's rating which will be received from IMDB score after it gets released in cinema. Thus, our solution will be more towards Supervised Learning Approach where we will be using imdb score in the dataset as the target variable (or response variable).

Movies having imdb score larger than 8.0 will be "**top movies**", and they are truly great movies from many perspective. Movies with rating from 6.0 to 8.0 are probably still "**good movies**". Viewers can gain something from them. Movies with rating from 1 to 5 are sometimes considered as ones that "**sucks**", in one way or the other. One should avoid those movies unless they have to. Life is short. Take a look at below diagram for general idea:



Our Approach

Model for the problem

Out of the 28 variables provided in dataset, we are especially interested in knowing how does the IMDB rating score correlate with other variables.

We will be using the Multiple Linear Regression/ Logistic Regression model approach to predict the imdb score. As we have to take a closer look on multiple predictors and their interaction between them which might impact the movie rating and to put the movie in one of our classified parameter, we have to use Logistic Regression approach.

Considered parameters in our problem

Before the release of a movie, it get lots of criticism based on director, actors (which are in cast), story plot and genre. They have their cool attractive facebook page before the release, provides certain great trailer(s) where the people provides their opinions/ reviews on that. We can get the movie popularity by the number of facebook likes, number of user providing their reviews, number of faces in movie poster and even by popularity of actors/ directors in the cast. Sometimes, directors/ producers even mentioned their overall budget for the film to indicate how much they are investing in to make it a big movie.

Below are some of the predictors we are planning to use from the dataset:

- Num_critic_for_reviews
- Director_facebook_likes
- Cast_total_facebook_likes
- Budget
- Num_user_for_reviews

We hope that based on the above and some more predictors, we will be able to come up with some great model to predict the greatness of the movie.

Dataset

We will be using the data (provided in csv) from the below mentioned kaggle website:
https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

Data is in 1.5MB having a set of around 5000 rows and 28 columns including the imdb_score which we will be using as target variable.

Training and Testing data will be split into the ratio of 80:20. Training data will be ~4000 rows.

Tools

We are planning to use R language and RStudio for our purpose.

Milestones

No	Item description	To do date	Status
1	Meet with team, decide on topic, features and target variable(s), split tasks	09/26	TBD
2	Download data, setup git repo with team, install R packages	09/30	TBD
3	Data exploration / preprocessing Project Proposal Due	10/03	TBD
4	Data preprocessing	10/10	TBD
5	Feature selection / creation	10/17	TBD
6	Run models, evaluate accuracy, plot accuracy results (Round 1)	10/24	TBD
7	Update models, run cross validation, evaluate accuracy, plot accuracy results (Round 2)	10/31	TBD
8	Round 3 (with significance of results always tested)	10/03	TBD
9	Project Progress Report Due	11/03	TBD
10	Plot comparisons: models (m1, m2, m3, ...), baselines (random), compute % improvements	11/07	TBD
11	Fix bugs, make improvements and import more new technics	11/24	TBD
12	Project Presentation	12/05	TBD
13	Project Report and Code Due	12/07	TBD