IMDB Data Analysis

Qingsong Wang

May 6th, 2019

1. Introduction

Movies, also known as films, are a type of visual communication which uses moving pictures and sound to tell stories or teach people something. Nowadays, movies are appealing a rapidly-increasing number of audience. There are abundant data resource, recording various kinds of features of thousands of movies. In this project, we aim at gaining comprehensive knowledge about movies, via analyzing related datasets. As listed below, three main goals are set here:

- Have a deep understanding about the dataset.
- Predict the IMDB score of a movie based on several variables.
- Determine whether a movie is profitable, or in other words, worth investing.

We want to achieve our goals by thorough explanatory data analysis as well as building popular machine learning models to predict or classify. Such models include support vector machine, random forest, neural network and etc. As a data analysis report, we try to show to the audience as many vivid results as possible, so the main focus lies on the the EDA part, while we still spend enough time on modelling.

The report is organized as follows: section 1 is a brief introduction; section 2 contains data description and pre-processing; we begin our explanatory analysis in section 3, where the most visualization results are shown; different models for prediction and classification are introduced and trained in section 4 with parameters tuned properly; we close this report with summary and discussion in section 5.

2. Data Description

After comparison, we decide to collect two datasets from Kaggle, and combine them. The reason why we choose these two is that they are complete and contain more intereting features. The combined dataset contains nearly 5000 movies as well as 30 features, and after removing NA rows 3636 pieces are left. The meanings of some variable are listed below, while the left are intuitive.

Variable	Description
aspect_ratio	ratio of length to width of screen
budget	the budget of the movie in USD
color	logical, whether is colorful or black-white
facenum_in_posters	number of faces in the poster
genres	movie type e.g. thriller, drama, etc
gross	the revenue of movies in USD
imdb_score	the audience rating
runtime	the duration of movies

For simplicity, from now on we only analyze 3494 movies in English, to remove the influence of language. And we add two columns, one named **if_profit**, which is a logical value indicating whether a movie is profitable, the other named **return_rate**, i.e. the ratio of net revenue to budget. Notice **return_rate** is more reasonable to measure the revenue, considering inflation. Then we can begin our formal analysis.

3. Explanatory Data Analysis