Lab 2: What Makes a Movie Successful? Fall 2021

w203: Statistics for Data Science

November 1, 2021

Introduction

What makes a movie "one of the greats"? While that question may elude both data analysts and artists for years to come, we are looking to find some effective indicators of box office success for films, in the hopes of maintaining a thriving movie industry for years to come. This project will focus on several commonly cited reasons for movie fiscal success and analyze how much of an impact each factor has.

The client for this analysis could be any major movie studio or its parent company, e.g. NBCUniversal or Walt Disney Studios. This is a significant issue, because box office sales decreased by 60% in 2020 during COVID-19 pandemic, leaving the movie theaters empty for months. This analysis will provide clues for what a studio can do to speed up the recovery and account for lost revenue.

The datasets we'll use for this analysis come from NBCUniversal directly, as well as several assembled lists from online polls (e.g. "Greatest Actors/Actresses of All Time" - Ranker.com).

We're planning on using profitability as the response variable for analysis in this study - using R to clean the empty and irrelevant cells, a "profitability ratio" can be determined for each movie based on its box office sales divided by budget (and controlling for foreign films and currencies). In assessing the value of name-brand recognition in Hollywood, we'll use R and its extensions to cross-check the presence of the "greatest" actors and actresses of all time, assigning a count to each movie to later plot against profitability. We'll be using an Ordinary Least Squares regression model to predict movie success as a factor of famous actor/director presence, online ratings, and other variables that we find significant.

Once we obtain an accurate model, we can obtain numbers such as the monetary value of a single point increase in rating, to explain the trends and interpret results. We can consider several situations accounting for premiere dates, return on investment (ROI), runtime along variables mentioned earlier. While there are certain limitations to this data set and analysis e.g. limited number of data points, some data entry errors, inaccurate labeling and biased focus on foreign films, and subjective descriptions of the movies, this should serve as a tool and starting point for aspiring and current filmmakers.

The products of this analysis will include the code and final datasets, as well as a written summary of the significant results.