Springboard Capstone Project #1 Proposal

Film Data Study

Using data to predict popular film composition

Introduction

Film is a multi billion dollar industry. Making a single movie in Hollywood costs around 70 to 150 million dollars. Production companies acquire the budget from investors and partner with various outsourcing companies such as CGI, motion capture and even agencies to create what the directors and writers had envisioned. A lot of the time, the production companies are able to earn what they have spent on making the movie. Many other times, the earnings are less than the cost. Very few occasions with just the right composition, a movie can bring in billions of dollars worldwide. So what is the "right composition"?

The problem

Every movie consists of thousands of different parts and factors. There are of course the actors and directors, producers and writers. And there are consultants, agents, crews like the cinematographers, sound engineer, lighting expert and even caterer, the list goes on.

To everyday movie goers, the public facing aspect of these "parts" are the main determinant whether average joe or ordinary jane decides that they might like the movie or not. By public facing parts, this will include the actors, directors, producers, production studio or even genre. And with around 600 movies being made every year, we see definite trends in a specific genre or film produced by a certain studio or director or all of those combined. For example, Marvel studios with Kevin Feige as executive producer making superhero films starring Robert Downey Jr. as Iron Man has been dominating the box office for almost a decade now. But as we all know, this wasn't the case 20 years ago.

The question we are trying to answer is: are there patterns or even cycles of genres that popularize? Are there combinations of movie “parts” that will produce a blockbuster?

This is for you

This analysis report will provide insights to any production companies or investors who are curious to know the industry’s past and current trends and who are interested in making data driven decisions for the future. This is for directors who are seeking for ideas on the next projects and for those who wish to reflect on their past films with respect to the market it was released on. Lastly this is for any actors, crews and enthusiasts who are looking to learn more about the film industry statistics in general and the way it has been changing so far.

The Data

Data used in this report will primarily be from Internet Movie Database also known as IMDb. The dataset contains about 10 million records of movie related information since 1892 and is refreshed daily with new updates. Data itself consists of basic information about the movie such as title, year released, language, mature rating as well as metadata like director, writer, star actors, runtime, characters and etc. Lastly, the data also contains user input data like rating of the movie and number of votes the movie has gotten for those ratings. The data is available to download for free from IMDb website : https://datasets.imdbws.com/ Information about the data can be found here : https://www.imdb.com/interfaces/

Trajectory

Throughout the report, the data will go through extensive transformations and analysis. Firstly, statistics of the data itself will be checked to make sure duplicates and unknown, null variables are taken care of. Actual data is some 7 different tsv files that share a common identifier for each movie. So identifying the feature to use and isolate, as well as append different datasets together to create one to few datasets to work with. Some categorical features may be translated to numerical variables and vice versa depending on the need.

After making sure the dataset is clean to work with, data will go through an exploratory data analysis phase to find out if there are any correlations between features, locate any outliers, and uncover interesting trends and patterns that lie within. Since the data contains a date feature, the dataset can be turned into a time series data to plot visualization of the change for the industry. This trajectory is subject to change as the curriculum goes on.

The Deliverables

By the end of this project, the deliverables will include the report itself, original dataset, transformed dataset, code used, and slide deck presentation of the findings.