Jupiter Analytics Recommendations for Microsoft Studios

This repo contains: Our collaborative Jupyter Notebook, which contains our code , the raw data files used in our analysis, and our presentation we made to deliver to the Microsoft Studio team

🔗 Analysis Goals

Our team was tasked by Microsoft to advise their fledgling studio on what types of movies would be the most successful.

🔗 The Data

Our data came from a number of sources, as no single database held all of the information required for our analysis.

IMDB - Internet Movie Database - The go-to website for information on all things movies, this served as the primary base for compiling the necessary information. IMDB's size was a double edged sword, its tables often consisting of tens of thousands of rows. This gave us a fantastic number of films and personnel to investigate, but sifting through that much data had its own challenges as well.

TheNumbers.com - For all of IMDB's greatness, it does not provide us with the financial information we wanted to add to our analysis. Thenumbers.com is a website wholly devoted to tracking budgets and revenues for films. The fact that it tracked both was critical, because that allowed us to calculate net profit by adding together foreign and domestic revenue, then subtracting the budget.

TMDB - The Movie Database - This database provided additional ratings and financial data

Box Office Mojo - This database was combined with the others to give additional ratings and popularity data

🔗 Our Methods

In order to make sure that our findings would be relevant to Microsoft setting up a studio in current times, we made the decision to only look at movies from the last ten years. This would ensure that our data did not point us towards trends from the past. If we looked at the entire history of film in our analysis, movies like Citizen Kane, or Gone With the Wind could have an outsized influence on our business recommendations.

Our team focused on net profit, and popularity of films as measures of success. As a metric, net profit allowed us to examine the most commercially successful films, telling us which genres, studios and cast and crew members have worked the best. In addition to this popularity allows us to take a less money centric approach. This is relevant to Microsoft Studios as box office revenues do not necessarily directly apply to streaming platforms. For example, popularity might be a better metric for advising Amazon Prime Video Movies, as the main goal of those projects is to bring more subscribers onto amazon prime as a whole. In this way popularity helps Microsoft Studios make sure that their film projects will bring subscribers onto Microsoft platforms.

Combining multiple tables both internally (joining together IMDB tables) as well as externally (integrating data from all four data sources) required a fair bit of exploratory data analysis, and clever problem solving. One specific challenge is due to the fact that IMDB is a global website, and hosts movies from all over the world. To avoid massive reduncies in their database storing every movie and person in every possible language, titles and names are instead stored as name and title codes, meaningless without the context of other tables. So one initial hurdle was simply joining together the tables from IMDB, translating name codes to names, and title codes to titles. This was key because getting a table that connected titles, to cast and crew, to their job on the film was essential for the cast and crew analysis. After that the IMDB main table ahd to be joined to the financial data from thenumbers.com. To achieve this the titles were stripped of all punctuation, whitespace, and all letters were put into lowercase for titles in both datasets. This was in an attempt to make the join go smoothly between two databases where the titles could have had small differences that would have lowered the sample size of our data. Once this main table was joined up however, it could be grouped by job title, grouped again by name, and the sum of net profits for each person's name could be found and ordered.

The resulting tables were then grouped and ordered according to each of our three analysis perspectives, and the results were visualized using a combination of matplotlib and seaborn.

🔗 Final Recommendations

Once we compiled all of the most successful studios, genres, and personnel, a clear trend emerged. The highest performing, most popular films come from well known franchises. Whether it be action and adventure coming out on top as the most popular genres, Disney crushing the competition based on revenue, or Kevin Feige, Rober Downey Junior, and the Russo brothers leading their job categories, it is clear that Microsoft should focus on acquiring beloved and well known intellectual properties. This could be achieved by merging with or purchasing an already exisiting studio like lionsgate, or the acquisition of the rights to well known IP's outright. Obviously Microsoft Studios will have to fill out its catalogue with films at all budget levels, but we believe the main draw for Microsoft's new streaming platform will be these large budget franchise films.

We also noticed the effect of biased hiring practices within the film industry, and we feel there is a huge opportunity for microsoft to lead the industry in fair hiring and pay practices. This combination of high budget action adventure films, adapted from well known intellectual properties, and a diverse cast and crew will set Microsoft Studios apart from its competitors and generate massive value for Microsoft platforms.