Microsoft Film Analysis Based On ROI

Author: Andre Layton

Summary

You're looking to break into the film industry, but stuck on what sort of content to begin creating. This study analyzes movie data from IMDB and financial data from The Numbers to recommend the best types of films to create for your studio's initial content.

The main focus behind the analysis is seeing which films are best to begin creating by calculating and comparing both the median Domestic and Worldwide returns on investment (ROI), and then analyzing production budgets to estimate how much investment will be necessary.

According to the analysis:

- Adventure films are the best genre to begin creating for high gross sales and ROI.
- A minimum budget of $100 million will be necessary to begin seeing profit.
- A maximum of $170 million to invest in the initial film in order to gross $200 million domestically.

Outline

- Business Problem
- Data
- Methods
- Results
- Conclusions

Business Problem

- Determine which film genre would be best to create for the new studio's initial content based on ROI data

- Estimate production costs

Data

- The dataset from The Numbers displays the budgets, domestic gross, worldwide gross, and release dates of various movies, ordered by budget in descending order.
    - The data from The Numbers will be crucial in calculating the ROI, which is equivalent to the gross income divided by the cost of the investment, or the budget in this case.

- The data located in IMDB is located in a SQLite database with various tables, one of which is "movie_basics", which contains the names, genres, and other details for each movie in the database.
    - Despite the need to calculate the ROI, the independent variable we are interested in is the movie genre and which genres yield the highest ROIs.

Methods

In order to combine both dataframes and create one single dataset for analysis, I begin by preparing the movie_budgets data first. I drop any fields I do not need (i.e., 'id') before I clean up the names of the remaining features.

I also change the data type of the last three columns, based on the information given by the .info() method, as they are numeric and key to statistical analysis later in our exploratory process. In order to convert these fields to integer types, the values for these fields must be transformed to fit conversion criterion. The .head() method also reveals that there are movie titles that contain apostrophes (') that did not fully convert, and thus requires further data munging.

Next step is to clean up the SQL table to fit for merging with our refined movie budgets dataset. Fortunately, this table doesn't require much work - only renaming the field names. Once that's complete, we merge the two together using the Movie Name field. After joining both tables, we begin cleaning up our final DataFrame by creating a new column representing whether the movie titles match to ensure the genres listed accurately describe the associated movie in each record.

Next, we find any missing values, and after taking a deeper look at the dataset, decide to remove these missing records as well as the records that do not have matching titles. Before dropping the records with mismatched movie titles, we create a function to change all movie names with numbers in the string to numeric form to bring a little more consistency in the movie names, and allow us to retain a few more records. We also drop all duplicates from the combined dataset to produce a final dataset for the next step in our process.

METHOD HERE FOR MODELING, ANALYSIS, SO ON BUT CONCISE

Results

Present the results of your analysis or modeling here. Should include evaluation of how well your results solve the business problem.

Conclusions

Present your conclusions about the project here. Can include business recommendations, project limitations, and/or future improvement ideas

Here are my recommendations based on my analysis:

1. **Focus on the genres with low margins of error.** While Musical and War movies have the highest returns on investment (ROIs), they are both in the bottom five in value counts, as well. As such, their margins of error are high - meaning that type of success may not be as consistent, or proven, as the plot shows. The genres with low error margins provide a better picture and potentially have more confidence in their results - primarily the top 7 with over 200 records.

2. **Stick with the top 3 genres.** Horror movies would not be a bad place to start; however, with the top 3 genres - and the majority of our data - being Drama, Action, and Comedy movies, I would focus my energy and resources on those genres that are popular in number and high in ROI, domestically and globally.

3. **Start with a Comedy or Drama, depending on your goals.** Comedy movies are shown to not only be popular in number, but also also produce high ROIs, and would be a great place to start. However, if the goal is to also get nominated for accolades and compete for awards, Dramas are generally the best route to go. Dramas, fortunately, fall in the top 3 in both ROIs, assuming we focus on the low error margins. In either route, I would also recommend casting high-profile actors/actresses to further promote the film and promote a high return on investment.

The analysis above, as mentioned earlier, is formulated assuming each movie has one individual genre that defines it. Unfortunately, films can be characterized in multiple ways/genres, and the analysis would need to be revise with that factored in, in order to solve the business problem when presented with those possible datasets. To further improve this project, I could also analyze ratings, see how that compares with ROI success, and observe any changes in results (or any confirmations). If I were Microsoft, I would also look into creating a streaming service and any research behind what other companies did or created to possibly factor that into the decision (i.e., Apple created AppleTV+ and produced a Drama movie for its first film, Amazon created Prime and produced a Drama as well, etc.).

Thank You!

Email: alaygt6@gmail.com
GitHub: @therookiescientist-andre
LinkedIn: linkedin.com/in/username/