Movie Box Office Success:

Factors Influencing Domestic Opening Weekend Revenue

**Github link: https://github.com/Ostendma/TOOL1\_FINAL\_PROJECT**

**Dataset and Motivation**

Our project focuses on domestic opening weekend box office revenue as a primary metric. This choice stems from the pivotal role this metric plays in gauging a movie’s immediate market impact and overall success. Opening weekend revenue is not just a reflection of the movie’s anticipation and marketing effectiveness but also sets a trajectory for its subsequent financial performance.

Our initial dataset was pulled from the “grouplens MovieLens” dataset (https://grouplens.org/datasets/movielens/25m/). The dataset contains about 62,000 movies from 1926 through 2023. This was the universe of movie entries we began with. The data we used from this initial dataset was the Title, imdbID and tmdbID. The latter two being unique identifiers used across various industry resources.

Using the imdbID from the initial dataset (“MovieLens”) we scraped the website: www.boxofficemojo.com to retrieve domestic opening weekend box office revenue (“opening revenue”) and the number of theaters a movie was released in. We decided to use domestic opening weekend box office revenue as it was a pure representation of how much revenue a movie generated whereas total revenue (“gross revenue”) includes metrics such as international sales, merchandise sales, streaming rights revenue, etc. From the roughly 62,000 movies, we only returned about 13,000 movie titles with the opening revenue and number of theaters. The remaining entries did not contain this information for various reasons; older movies that lacked this data set.

From the remaining dataset of roughly 13,000, we were able to use the imdbID once again to connect to the “OMDB API” (www.omdbapi.com) to retrieve the following information:

* Title of the movie
* Year of release
* Rating (MPAA rating. Ie; PG-13, R, etc)
* Release Date
* Runtime
* Genre
* Director
* Writer
* Actors
* Plot
* Language
* Country of Origin
* Awards
* IMDB Review Rating
* Rotten Tomatoes Review Rating
* Metacritic Review Rating
* Metascore
* Imdb Votes

We then connected to The Movie Database (www.themoviedb.org or “TMDB”) to retrieve a movie’s budget, overview, production company and collection a movie belongs to (such as “The Transporter Collection” or “Harry Potter Collection”). We also did a scrape of TMDB for a movie’s “featured review” which is a review that is randomly chosen when reaching a movie’s information page on TMDB. We would later use this to conduct an NLP Model regarding review sentiment.

Since the movies we analyzed spanned across several decades, we would also need to adjust opening revenue for inflation. We retrieved monthly CPI data from “https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/” which spans from January 1913 to September 2023.

**Actual Task Definition and Research Question**

**Real-World Problem**

In the current era marked by the surge in popularity of streaming services and the increasing affordability of home entertainment systems, the film industry faces unprecedented challenges. With an abundance of choices at their fingertips, consumers are more selective than ever. This evolving landscape has intensified the competition among movie studios to capture audiences' attention. Accurately forecasting a film's performance upon its release has thus become a critical task for these studios. Our project addresses this need by aiming to predict the opening weekend box office revenue of movies in the U.S. market.

**Research Question**

Our research question is twofold: "What are the key factors that predict a movie's opening weekend box office revenue in the U.S.?" and "How can these predictors be effectively modeled to forecast a movie’s performance?" This exploration is crucial for movie studios to allocate resources effectively, strategize marketing efforts, and make informed distribution decisions.

**Input of the Analysis**

The input for our analysis is a comprehensive dataset that encapsulates a broad spectrum of movie attributes. This includes:

* Opening Weekend Box Office Revenue
* Number of theaters at release
* Budget
* The collection a movie belonged to
* Year & Month of release
* MPAA Rating (R, PG-13, PG, G)
* Runtime
* Genre
* IMDB ratings
* Overview and review sentiment

**Output of the Analysis**

Our analytical endeavor culminates in a Random Forrest Regressor model designed to estimate the U.S. opening revenue of movies. The number of theaters a movie was released in and its budget end up being the most influential features to predict opening revenue. This model is intended to provide a tool that blends various movie attributes to yield an accurate forecast of a film's initial financial performance. It aims to serve as a model for movie studios and industry professionals in navigating the highly competitive and rapidly evolving film market.

**Literature Review**

There are certainly a handful of similar studies on the topic of predictors of movie economic success. In fact, the National Library of Medicine did research on what factors are important for movie financial success by applying a random forest analysis. They looked at technological changes, profits free of inflation, and used a smaller set of features. Their datasets included all films with budgets disclosed at the Box Office Mojo website, which resulted in in 3,167 movies released in theaters worldwide from 1980-2019.

Another study we found that is like the work we are doing was done by Durges Samariya. His analysis on box office revenue works on TMDB box office prediction datasets available on Kaggle. His approach to this analysis is broader. He investigates things like *Which movie made the highest revenue? Which movie has the highest budget?, Which movie is the longest movie?* and *In which year were most movies released?*, to name a few. Though the questions he answers in his research are interesting, he doesn’t go into depth about the drivers of revenue and whether there are features that can help predict revenue.

Our research stands out in several aspects. Initially, we concentrate on domestic films, specifically targeting their opening weekend revenue. Our thought behind this approach was to attain a more complete dataset. This would be beneficial as it would facilitate a more effective analysis to identify features that had substantial influences on predicting revenue. Another unique angle we took was to look at number of theaters a movie was released in. While it is commonly assumed that a greater number of theaters would result in higher revenue, we aimed to validate this assumption using a random forest regressor. Additionally, we enhance the exploratory aspect by incorporating sentiment analysis into the study. We found it intriguing to investigate whether the content of a film's overview (plot) or review influences people's decision to watch a movie, ultimately impacting revenue.

**Quality of cleaning**

As mentioned in the initial data collection, a large portion of the data did not have the primary metric: opening weekend revenue. Roughly 13,000 entries did have this data and became our primary dataset.

After loading our dataset into a Pandas DataFrame, we found a significant number of null values in the variables: Writer, Awards, DVD, Production, Website and Belongs\_to\_Collection. We removed the first 5 variables from our analysis but kept the collection data as we planned to feature engineer this data. Since we are looking to predict US domestic box office opening weekend revenue, we dropped films that were not released in the US.

The factors: Opening Revenue, Number of Theaters, Runtime, Ratings\_IMDB, Ratings\_Rotten and Ratings\_Metacritic were transformed to be the correct data types. Opening revenue and runtime were transformed into a float, number of theaters to integer, and the ratings scores were transformed to percentages. Since the movies span across several decades, we adjusted the revenue figures for inflation using monthly CPI data. We created a month and year column to match the corresponding CPI monthly figures.

After making these adjustments, we ran a correlation matrix and found that Ratings\_IMDB, Ratings\_Rotten and Ratings\_Metric were highly correlated to one another. This is not a surprise since they are all similar rating systems. Since Ratings\_IMDB had the highest correlation to CPI adjusted opening weekend revenue, we dropped the other two factors to improve model efficiency.

We then examined the remaining null values and found that the ‘belongs\_to\_collection’ accounted for most null values. We will go on to feature engineer this variable. We proceeded to drop all entries with null values not in the collection variable. We then transformed the Released Year and Month columns to be integers.

**Exploratory Data Analysis**

We first looked at the number of movies released by month to detect any seasonality. We did not find significant seasonality but rather a consistent release of movies each month.

A graph of blue bars

Description automatically generated

We then looked at the data over time; the number of movies released by year and the average number of theaters movies were released in by year.

A graph of a number of sales

Description automatically generated with medium confidence

A graph of a number of people

Description automatically generated with medium confidence

As time has passed, we have seen an increase in the number of movies released each year. Our dataset did not capture all movies released after 2019, therefore we did not put any weight on time. These two graphs show us when the movie theater industry itself became bigger. Around 1979 is when we see the start of an increase in movies being released each year. This also roughly lines up with the average number of movie theaters a movie was released in as see this uptick begin in 1976. When we compare the number of movies released from 1976 – 1978, we don’t see a large number of movies released but this is when films had wide releases to several movie theaters. Therefore, we removed any movies released prior to 1976.

We also looked at the distribution of the MPAA ratings. Since we are only looking at movies we removed any rating that is not part of the US movie rating system. Therefore, we only kept entries with ratings G, PG, PG-13 and R.

A graph with blue squares and black text

Description automatically generated

For the ‘belongs\_to\_collection’ we decided to feature engineer this column. The collection is in reference to whether a movie had a certain brand attached to it such as “The Harry Potter collection.” With 699 unique collections, instead of dummy coding each collection we created a binary Yes vs No (being in a collection) to see if such classification would impact our model. Note that the column ‘Released Year’ in the table below is the variable, whereas the numbers represent a count. (ie; 7062 films were not in a collection whereas 1330 were)

A black screen with white text

Description automatically generated

Doing further data cleaning and exploration, we found entries with a budget of $0 (instead of a null value) which were subsequently dropped. Regarding genre, films may be associated with several genres. Therefore, we coded the genres to be True/False for each movie. We then dropped all other unneeded columns from our dataframe that would not be applicable to predicting opening revenue whether it be columns such as URL or columns that were feature engineered or adjusted. Below is the final correlation matrix of the factors we included in our model. We see a high correlation between revenue and the number of theaters at release and the budget as well.

A screenshot of a graph

Description automatically generated

We then scaled our data to normalize and prevent any one feature from becoming too important in our models. We also dummy coded the categorical variables as a necessary step to run our models. We split the data into 80% training and 20% test sets resulting in 4,358 training entries and 1,090 test entries.

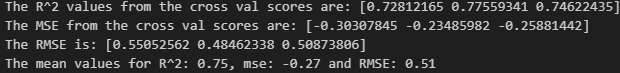
**Model Analysis**

We ran the following models on the training set: Linear Regression, Decision Tree Regression and Random Forest Regression. Of the 3 models we found that the Random Forest produced the best results.

A black screen with white text

Description automatically generated

We then did a grid search to find the best hyperparameters for a Random Forest model. It found that the best parameters were a max\_depth=40, max\_features=’sqrt’ and n\_estimators=1200 and bootstrap=False. The mean values for R^2, MSE and RMSE from the best parameters were:



The discrepancy in values from the initial Random Forest model to the model with the best estimators from a grid search could be that the initial model was overfitting the data.

The top 10 important features from the model resulting from the grid search were as follows:

A graph with numbers and a bar

Description automatically generated

The results of our model show that the most important variables that predict opening revenue were the number of theaters at release and the budget. This intuitively makes sense as a wider release allows for a film to capture more ticket sales through sheer volume. The budget is also indicative of how much marketing spend a production studio had available to advertising and promotion of a movie prior to release.

**NLP and Further Analysis**

To refine our analysis further, we ran a movie’s overview and featured review through a Natural Language Processing model to obtain a sentiment score. Many of the movies did not have a featured review that we could put through our NLP model. As a result, the resulting data frame was 4,664 entries.

To conduct the NLP Sentiment Analysis, we first read in the data from the ‘OMDB API Data’ and ‘Featured\_Reviews\_with\_tmdbdid’ csv files. We then convert them to individual pandas data frames. Once in pandas data frame, we create a function called ‘stopwords\_lemmetize’ to remove any stop words and lemmatize remaining words from the overview and featured review columns and create new columns with the results. After this process another function is created called ‘sentiment\_score’. This function runs the SentimentIntensityAnlyzer to analyze the columns ‘review\_NLG’ and ‘overview\_NLG’ and extracts the ‘compound’ score from the results. Finally, through an inner join, we join the two data frames together to remove any movies that did not have a featured review.

Below is the correlation matrix after adding the overview and review sentiment scores to the data frame.

A colorful squares with white text

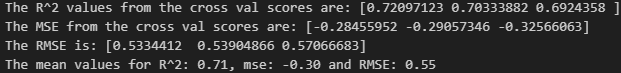
Description automatically generated

After scaling our data, re-adding the dummy-coding for categorical variables, and doing a train-test-split of 80-20, we ran this data through the same process as our previous models. The Random Forest Regression returned the best results.

A screenshot of a computer

Description automatically generated

We then did a grid search including the NLP data to find the best hyperparameters for a Random Forest model. It found that the best parameters were a max\_depth=40, max\_features=’sqrt’ and n\_estimators=1200 and bootstrap=False. The mean values for R^2, MSE and RMSE from the best parameters were:



The top 10 important features from the model resulting from the grid search were as follows:

A graph with numbers and a bar

Description automatically generated

Comparing the two models, we see that the model that doesn’t include NLP performs slightly better than the model that includes it. The model without NLP produced an R^2 of 0.75 whereas the model including NLP produced an R^2 of 0.71.

**Test Set**

We ran the Random Forest Regression model excluding the NLP data against our test set. It produced an R^2 of 0.72 and a RMSE of 0.46. This model results in a robust model to predict opening revenue as it captures 72% of the variability.

Below is the top 2 splits of the tree. Due to the depth of the tree, we could not display the entire tree in a readable output.

A diagram of a triangle

Description automatically generated