The Economics of Film: An Investigation of the Financial Interplay Between Film Releases and Stocks Market Trends

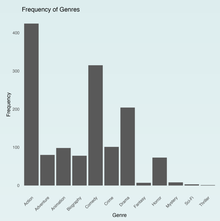
Lizzie Healy, Rachna Rawalpally, Sophia Rutman, Amina Nsanza
Georgetown University Data Science and Analytics

Introduction

The impact of the film industry on the broader economy has been a significant area of research since films first became a popular pastime. Conversely, films have also served as a reflection of societal conditions, particularly since the post-war era. This work focuses on identifying this bidirectional connection between the state of the economy and the film industry.

Specifically, we will investigate how film releases impact the state of the economy and, in turn, how economic conditions influence the funding of film production. This approach allows us to better understand the interconnectedness between the two entities and account for the extent to which they are inherently linked.

We utilize the stock market closing values, particularly, the S&P index to proxy for the strength of the economy and box office values to measure film success. To begin, we will employ feature selection on both the film and stock variables. This will be followed by a series of regression models investigating both directions of research and iteratively including variables to increase the robustness of this research.



Data

Our data ecompasses 20 years of movie releases and S&P 500 data, with a total of 4000 films and daily information of 35 separate publicly traded corporations. Our features are presented below:

Film Features:	Financial Features:
Genre, Runtime, Rated, IMDB Rating, Metascore, IMDb Votes, Box Office, Country, Language	Open, High, Low, Close, Adj Close, Volume, Ticker

The genre, rating, country, language, and ticker were encoded to ensure numerical data throughout the set. As a basic overview of Internet Movie Database (IMDb) terminology, the IMDb Rating is an averaged rating across all IMDb users that engaged with a particular film, and IMDB votes is the total number of viewers that engaged with the film in general.

As for the financial features, open represents the first price of the day for one share, and close represents the last while the market is open. The high and low are the largest and smallest prices on the day, and volume is the total amount of shares bought or sold. Adjusted close is the stock's closing price after accounting for corporate actions like dividends, stock splits, and rights offerings. It gives a more accurate reflection of the stock's value over time for analyzing historical performance.

We joined these two datasets on the release date of a particular film. All 35 stocks are associated with a film given the stock information is available on that day and a film was released on that day.

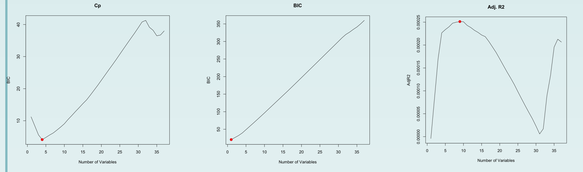
Feature Selection

In order to decide on which stocks were most important in our model, we ran forward, backward, and complete subset selection using the *regsubsets* function in the *leaps* R package.

Finance

We hypothesized that entertainment corporations would be most influential in a model that uses purely S&P 500 data to predict box office revenues for a film, but found this to be false.

The forward and full feature selection were in agreement with features that optimized Mallow's Cp (Cp), Bayesian Information Criterion (BIC), and Adjusted R^2 values. The backwards feature selection was not, so we chose to ignore these findings seeing as the full feature selection is more comprehensive. However, these three criterion dictated that the optimal number of features were 4, 1, and 9, respectively. The Cp selected Open, High, Adjusted Close, and SPX, BIC selected Adjusted Close, and Adjusted R^2 selected SPX, XOM, T, GOOG, LDOS, Open, High, and Adjusted Close.

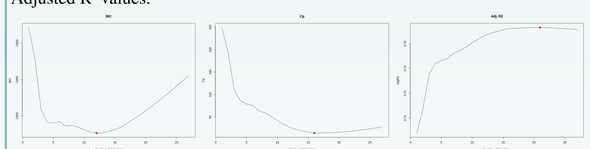


We hypothesize that SPX, XOM, T, GOOG, and LDOS were selected because they indicate general economic trends just as box office tallies often do. SPX is a general index for 500 large corporations, so this will fluctuate with these corporations. T and GOOG are large tech companies that will gain and lose revenue through economic cycles, as will XOM as a giant in the energy sector. Finally, LDOS tracks government spending, which may correlate with public behavior. Due to these reasons, we use these indexes in our modeling.

Film

For the film dataset, we attempted to discover the characteristics of a film that were most influential on the S&P 500 index.

The results across the full, backward, and forward selection were in slight disagreement as was the additional measure of lasso selection. Backward and full feature selection both selected the same variables across BIC, CP, and Adjusted R^2 values.

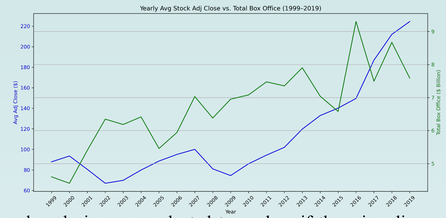


However, these yielded 12, 16, and 21 features respectively. Forward produced a model with 5 features based on BIC and 21 features based on CP and Adjusted R^2 values. Furthermore, Lasso selected just five features.

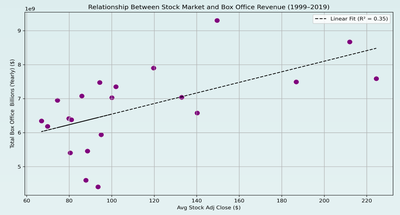
The only four variables that were constant across the entire analysis were Year, Runtime, Metascore, and IMDb.Votes. The additional variables were different choices of the MPAA rating and the genre of the film. Moving forward we will include these four core variables as our baseline model and iteratively include the genre and rating one-hot encoded variables that were present in the backward and full subset selection models to discuss which improve the predictive power in this scenario.

Modeling - Film to Finance

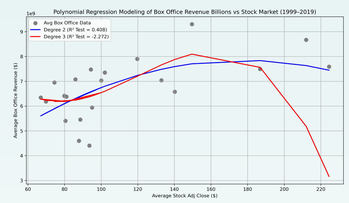
To explore the relationship between the stock market and box office, we analyzed the relationship between the average adjusted stock and total yearly box office profits between the years 1999 - 2019. From the initial plot we observed that there is no consistent yearly alignment, that there is a volatility that the box office experiences that the stock market does not.

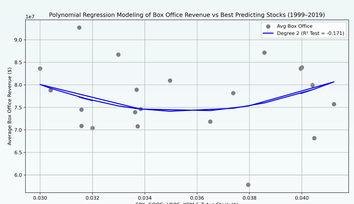


An additional analysis was conducted to explore if there is a linear relationship present between stocks and box office prices. Using a simple linear regression model, a slight positive trend was observed with an R^2 of 0.35, indicating that only 35% of the variation in box office can be linearly explained by stocks.



Building on the analysis above, a polynomial regression model was selected to predict box office prices; due to its ability to capture potential nonlinear patterns. The baseline model tested 2 polynomial degs, 2 and 3. It was observed that that the lower deg 2 outperformed the deg 3 with an R^2 of 0.408.

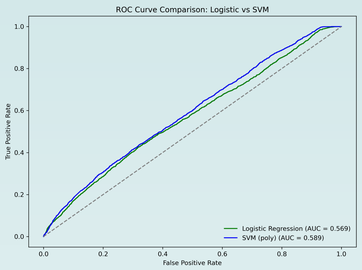




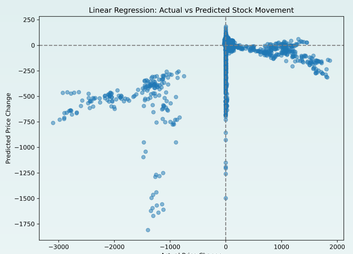
The baseline model was refined using the stocks identified in the feature selection. However, the new model did not perform as expected and yielded an R^2 of - 0.171. This weakened the model's performance and suggests that polynomial modeling on a reduced dataset may not always provide the best results. Furthermore, box office performance could not be predicted simply by the stock market, suggesting that there are additional factors beyond economic variables that affect movie revenues.

Modeling - Finance to Film

We used Support Vector Machine (SVM) and Logistic Regression to predict whether the stock price would go up or down the next day based on movie features. To enhance predictive power, we also added lagged stock variables (ex. yesterday's price, and return), which help capture recent market momentum or trends that might influence future price direction. The ROC curve shows that both models performed similarly, with AUC scores ranging from 0.569 to 0.589.



We applied linear regression to predict the magnitude and direction of stock price change using movie features and lagged stock variables. The model produced an R^2 of 0.17, explaining only about 17% of the variance in stock price movement. The RMSE of 290 and MAE of 110 also reflect significant average errors, especially relative to typical daily price fluctuations.



Conclusion

Across both classification and regression models, we found that movie-related features, even when combined with lagged stock variables, did a poor job of predicting the stock market. Model performance remained weak despite using a range of modeling techniques (logistic regression, SVM, linear regression), along with regularization, and cross-validation. Also, adding more features (e.g. genres, rating, runtime) did not improve model performance. These results highlight that the stock market is highly volatile and influenced by complex factors far beyond movie metadata. These results were consistent with the prediction of a film's box office revenue for the polynomial regression, suggesting that looking at economic trends alone is not enough to predict box office revenues.

In summary, movie data and short-term stock movement alone, cannot independently serve as reliable predictors for each other. Our models' performance proved this by yielding lower R^2 and higher RMSE & MAE, implying that additional variables are needed to effectively predict economic trend and movie industry profits.

References

We would like to express my gratitude to Professor David Byrd of Bowdoin College for his work in compiling the time series data on the S&P 500.

Our additional data collection was sourced from OMBd API for movie features.