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

Lizzie Healy
Rachna Rawalpally
Sophia Rutman
Amina Nsanza

DSAN 5300
Georgetown University
April 30th, 2025

ABSTRACT

This paper investigates the financial interplay between the film industry and the stock market, examining how movie releases may influence market trends and vice versa. Using a dataset combining stock performance data from major S&P 500 corporations and movie characteristics scraped from IMDb via the OMDb API; we conducted regression analyses, feature selection, and predictive modeling to explore these relationships. We evaluated both directions of influence: how stock trends relate to box office revenues and whether movie features can predict stock market movements. While binary classification models such as logistic regression and SVM performed poorly, suggesting that movie features alone cannot predict stock movements, more complex models like XGBoost captured some nonlinear patterns, achieving an R^2 of 0.76 when predicting stock changes based on selected features. Overall, the study highlights a limited connection between economic indicators and film success, suggesting avenues for future research using more advanced time-dependent models.

Section I. 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. Understanding how the film industry operates, in itself, is a profitable business to the tune of over 92 billion dollars annually (Statista 2024). Industry giants and cinephiles alike have endeavored to create the perfect formula for film success, piecing together cinematography styles with groupings of actors in attempts to turn out blockbusters. However, success can not be manufactured, especially when it hinges on the erratic and inconsistent nature of the general public.

Similarly, attempting to predict stock market trends is even more lucrative, boosting a market capitalization of nearly 32 trillion dollars (Statista 2025). While a worthwhile research endeavor, many have found that the stock market is almost unexploitable with its complex interdependencies and extreme volatility. To the dismay of many, the precise movement of the stock market is yet beyond our reach, however, some have found correlations between market trends and external factors, small indicators that offer insight into what drives the stock market causing it to tick in any given direction on any given day.

One of these external factors is films, which have served as a reflection of societal conditions and, by proxy, the stock market. This trend emerged in the post-war era when the gloomy nature of the nation was changing the sentiment of film production and the film industry was seeing changes to attendance and ticket sales, impacting the larger economy. This connection, while apparent, is often observed ex-post, meaning it has not, as of yet, been the secret to unlocking the mysteries of the stock market. It does, however, illuminate that films often mirror economic sentiment and that they may offer subtle clues to the forces behind the financial markets.

This work focuses on identifying this bidirectional connection between 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 film revenues. This approach allows us to 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. Our unit of analysis is individual days marked with a particular film release and complementary stock data.

We hypothesize that an underlying relationship between the film industry and stock market movement exists and is observable through statistical analysis.

Section II. Literature Review

A comprehensive review by Chhajer et al. (2022) highlights the growing role of artificial intelligence, specifically artificial neural networks (ANNs), support vector machines (SVMs),

and other deep learning models, in predicting stock market behavior. Their study emphasizes that traditional linear models such as simple regressions often fail to capture the complex, nonlinear patterns inherent in financial data, motivating the transition toward more flexible machine learning frameworks. Artificial neural networks, with their ability to model complex relationships between inputs and outputs without requiring explicit programming, have been extensively applied in stock market forecasting tasks. SVMs maximize the margin between data points and decision boundaries, avoiding overfitting as long as extensive hyperparameter tuning is completed. Deep learning models, including convolutional and recurrent neural networks, have emerged as a powerful extension of traditional ANNs. The review notes, however, that deep learning models require large datasets and significant computational resources, which can be limiting factors for their widespread adoption in some financial settings. Despite these advances, Chhajer et al. also point out several ongoing challenges. Overfitting remains a concern, particularly when models are trained on small or unbalanced datasets. Furthermore, financial data is highly volatile and often influenced by exogenous factors that machine learning models may struggle to capture. The lack of interpretability in complex models like deep neural networks also presents barriers to trust and regulatory acceptance in financial industries. Due to these issues, we mainly focus on more flexible versions of linear regression, such as polyfit models, SVMs, and some basic deep learning algorithms such as XGBoost (Chhajer, Shah, & Kshirsagar, 2022).

Shifting from stock markets to the film industry, another line of research applied to predictive frameworks to understand box office performance. In one particular paper, the authors aimed to determine the factors that affect box-office revenue by looking at how content positively or negatively affects the financial success of a film. Specifically, they utilized films from the late nineties to the early two- thousands to look at how the quality of the movie, as measured by the cast's critical acclaim and the budget, combined with the offensive content, as measured by R-ratings, are a major factor in the revenue generation of a film. In the empirical analysis, box office revenue, adjusted for inflation, was used as the response variable and the predictors were film quality (vector of earnings and awards earned by the cast), MPAA rating, film budget, cast talent, the distributor of the film, award success, the season of film release, and the offensive content intensity. Utilizing ordinary least squares regressions, they achieved results that indicate movie quality is a major factor in the revenue generation of a film. Specifically, adventure, comedy, horror, and romance have the highest effect on revenue, while Westerns have the most negative effect on revenue. In addition, violence attracts a greater audience and more revenue. Seasonally, releasing a film during the summer has a positive impact on revenues, and releasing it in the fall has a negative effect. Furthermore, the inclusion of offensive content specifically profanity, sex, and nudity hurts box office revenues. In conclusion, quality is of strong importance to the success of a film and offensive content may not be well-perceived by audiences (Garcia-del-Barrio & Zarco, 2016).

Complementing the film research, Einav and Ravid explore the stock market's response to movie opening date changes. Their main goal is to understand how announcements about

release date changes influence stock price movements and what factors drive these reactions. They find that the stock market reacts negatively to announcements of release date changes, regardless of whether the new date is specific or vague. This parallels findings from broader finance research, where the market tends to react negatively to revisions in corporate information. Focusing on the film industry, they observe that movies with higher production costs experience larger adverse stock market reactions, suggesting that investors are more concerned with budget risks than potential revenue gains. Interestingly, they also find that market reactions are not significantly correlated with a movie's actual box office performance, indicating a limited ability of investors to predict future success based solely on release date changes. Their study, based on 302 changes involving 260 movies from 1985 to 1999, uses event study methods with a window of five days before and after announcements. Overall, the paper highlights that even small events, like a movie delay, can impact stock prices meaningfully (Einav & Ravid, 2009).

Taken together, these papers highlight the current areas of research in predictive modeling, whether applied to financial markets or the entertainment industry. They highlight the potential value of forecasting, while also revealing the challenges and limitations of the current research landscape.

Section III. Data

Our stock market dataset encompassed one CSV file for each S&P500 corporation, where the unit of observation was one day. It also included an S&P index. The features were opening price, closing price, volume traded, adjusted close, and date. The opening price was the price when the stock market opened on a particular day, and the closing price was the exact price at closing. The adjusted close, however, takes into account dividends and splits. It gives a more holistic view of the stock price over time because this number is not affected by those types of price increases or decreases. Volume is the number of shares bought and sold on that particular day. The dates vary for each CSV but generally extend from the late eighties and early nineties to 2023.

To combine all of these CSVs, we selected 32 indices to use within our project. These indices include large companies that we felt increased and decreased with the economy over time, and corporations specifically in the technology sector. Our final list included: AAPL (Apple Inc.), BBY (Best Buy Co., Inc.), BIO (Bio-Rad Laboratories, Inc.), CERN (Cerner Corporation), CMCSA (Comcast Corporation), DIS (The Walt Disney Company), DISCA (Discovery, Inc.), ENPH (Enphase Energy, Inc.), FCNCA (First Citizens BancShares, Inc.), GOOG (Alphabet Inc.), HUBB (Hubbell Incorporated), JBL (Jabil Inc.), KEYS (Keysight Technologies, Inc.), LDOS (Leidos Holdings, Inc.), MASI (Masimo Corporation), NATI (National Instruments Corporation), NFLX (Netflix, Inc.), ORCL (Oracle Corporation), PFE (Pfizer Inc.), QCOM (Qualcomm Incorporated), RGEN (Repligen Corporation), SAGE (Sage Therapeutics, Inc.), T (AT&T Inc.), TECH (Bio-Techne Corporation), UBER (Uber Technologies, Inc.), VIAC (ViacomCBS Inc.), VOYA (Voya Financial, Inc.), WIX (Wix.com

Ltd.), XOM (Exxon Mobil Corporation), YETI (YETI Holdings, Inc.), ZION (Zions Bancorporation), and SPX (S&P 500 Index). We merged these CSVs by concatenating them and adding a column to encode the ticker. The data was repurposed from Professor David Byrd's Financial Machine Learning course at Bowdoin College.

To obtain all the necessary movie features, we used the OMDb API to access information from IMDb. We repurposed a movie dataset from a previous project in DSAN 5400, using the list of movies as the basis for our data collection. Through the OMDb API, the resulting film dataset contains detailed information on 4,277 films.

For the purposes of this analysis, the variables used were Year, Genre, Runtime, Rated, IMDb Rating, Metascore, IMDB Votes, and Box Office. The genre variable indicates all genres of the film, however, we will solely use the main genre, leaving us with 11 genres. The runtime indicates the length of the film in minutes. The rating variable specifies the Motion Picture Association of America (MPAA) rating of the film, which describes what age group the film is suitable for (ex. G, PG, PG-13, R). The IMDb rating is a score given by the Internet Movie Database, which takes scores given by IMDb users and aggregates them to generate a single rating from 0 to 10 for each film (IMDb 2025). The ratings in this dataset range from 1.5 to 9.3, with a 10 indicating a perfect film. The IMDb metascore is a similar metric that rates a film utilizing weighted averages of movie reviews from a select number of top movie critics (IMDb 2014). These range from 1 to 100, with 100 indicating the highest revered film according to critics. IMDb votes are the number of ratings the film has received from users, which correlates to the number of rating values used in the IMDb rating calculation, ranging from 5 to 3,015,278 votes in this particular dataset. Finally, the box office number is the value of tickets sold for the film's release in theaters, which is often used as a proxy for revenue and success of the film.

In order to combine the stock market and movie observations, we joined the two datasets on date. We completed an inner join to ensure that all of the observations we accrued could be used in some predictive capacity by the models we created. We filtered these dates to ensure our data set only included dates between 1999 and 2019, to avoid any abnormalities surrounding COVID-19, and to ensure that each movie observation was joined with every stock index upon joining.

Section IV. Methodology

A. Feature Selection

Feature Selection is a statistical method that finds the best subset of variables to utilize in a regression model. In other words, it determines which features collectively have the most impact on the response variable. The purpose of feature selection is to reduce the size of the model, increase efficiency, and prevent overfitting. We will employ the three main feature selection wrapper methods: best, forward, and backward subset selection. Each of these follows the same basic process of searching through different combinations of variables and selecting the one that optimizes some predefined criterion, with small changes in the process.

Best subset selection is an exhaustive search that looks through every combination of regression equations, however, it is time-consuming and subject to overfitting. Forward stepwise selection starts with an empty regression equation and adds the variable that gives the greatest improvement to the performance metric (ie. the residual sum of squares) at each step. Backward stepwise selection works in reverse order, starting with all of the variables and removing the least important one in terms of its impact on the metric at each step. These are both more efficient, but run the risk of not finding the optimal subset. We will also employ Lasso selection, which shrinks some unnecessary variable coefficients to zero, essentially performing feature selection, as a secondary check.

We will utilize Regsubsets from the *Leaps* R package for the three wrapper feature selection methods and R's *Glmnet* for Lasso Selection.

To choose which stocks were most important in our analysis, we completed feature selection in accordance with our explanation above. We then one-hot encoded the stock tickers before passing the data into *Leaps* forward, backward, and best feature selection. This led to a total of 37 features for the selection model to choose between. The models were told to predict box office revenue just using financial information. While these models were not powerful, they did provide an insight into which features are most relevant for our modeling.

All three methods provided the same results for optimal features in terms of BIC, Cp, and Adjusted R^2, so below we show the visuals collected from the best feature selection.

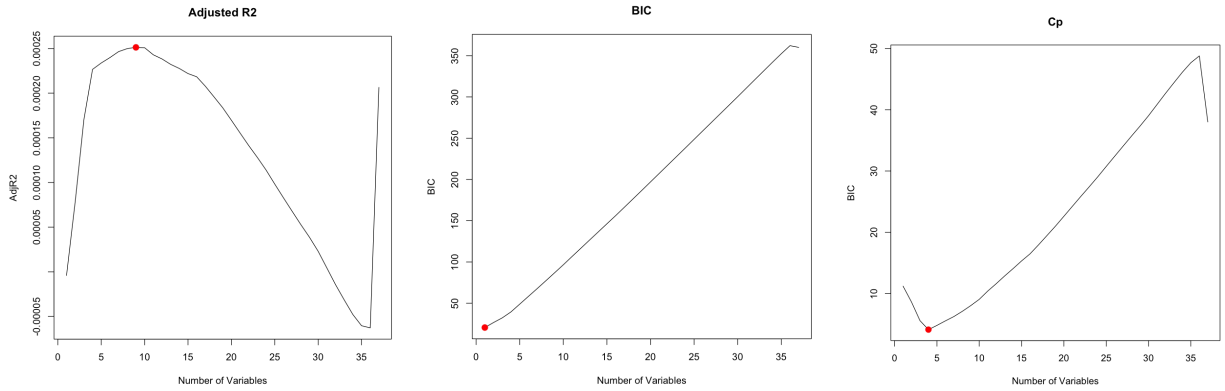


Figure 1. Optimal Adjusted R^2, BIC, and Cp values for stock feature selection

Figure 1 shows a discrepancy in the optimal number of features. Adjusted R^2 indicates that nine features are optimal, BIC suggests one, and Cp suggests four. 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.

The fact that adjusted close is included in every subset shows its importance in financial forecasting. By providing an insight into the pricing that is not affected by a particular corporation's decisions regarding shareholders, it becomes very powerful in predictive models. We decided to move forward with the nine features selected by the Adjusted R^2. 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 indices in our modeling.

We also completed a Lasso Regression to reaffirm our findings from above. However, this was not the end result. Figure 2 demonstrates the regression.

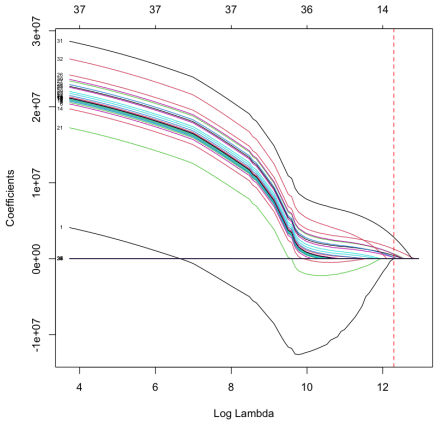


Figure 2. Results of Stock Lasso Feature Selection indicating each of the coefficient paths as a function of the logarithmic of lambda values.

However, this feature selection did not match that of the *Leaps* package. The final nine non-zero coefficients displayed were Adjusted Close, Low, High, Open, YETI (YETI Holdings, Inc.), KEYS (Keysight Technologies, Inc.), WIX (Wix.com Ltd.), VOYA (Voya Financial, Inc.), MAS (Masco Corporation), and LDOS (Leidos Holdings, Inc.). Although this reaffirmed the importance of the open, close, and adjusted close prices, it did not select any of the stocks identified by the previous subset selection.

We decided, despite this, to use the features chosen by the *Leaps* package. These features were verified by three different techniques, including a very comprehensive full feature selection algorithm. In addition, the Lasso algorithm is notorious for eliminating features that correlate with one another, even if both are necessary in maximizing predictive ability. The best subset selection looks at the best model for every set of features to avoid this problem and focuses on the best output instead of dependencies.

To perform feature selection with the film features, we merged the film characteristics and stock datasets. From here, we excluded any dates that did not coincide with a movie release or without pertinent stock data. In addition, this analysis solely used the Adjusted Close (Adj.Close) variable for the S&P 500 index stock; all other stocks and stock metrics were dropped before performing the selection. The poster and movie description variables from the

film side of the data were dropped as well as any missing values and NA values throughout, lending to 1,392 total observations.

We then conducted the necessary data preprocessing. Specifically, we one-hot encoded each of the categorical variables and ensured that the continuous variables were correctly formatted as numeric. The genre category was simplified to include only the main genre of the movie.

With the dataset prepared, we began feature selection to discover the characteristics of a film that were most influential on the S&P 500 index. For the criterion to choose a model, we utilized the Bayesian Information Criterion (BIC), Mallow's Cp, and Adjusted R^2.

	BIC	**CP**	**Adj R^2**	**Lasso**
Best	12	16	21	-
Forward	5	21	21	-
Backward	12	16	21	-
Lasso	-	-	-	5

Table 1. The number of predictors chosen by each type of stepwise feature selection and lasso methods based on the three criteria, BIC, CP, Adj R-Squared.

As depicted in Table 1, the results across the best, backward, forward, and lasso selection were in slight disagreement. Best subset and backward stepwise feature selection were in agreement on the number of variables to include as well as which variables across BIC, Cp, and Adjusted R^2. However, these yielded different regression results that included 12, 16, and 21 features, respectively. Forward stepwise, however, contradicted these results with regression sizes of 5 and 21 for BIC and Cp, and only had a matching result in terms of Adjusted R^2 with 21 features in that regression set. This regression set included the features: Year, GenreAdventure, GenreAnimation, GenreComedy, GenreCrime, GenreDrama, GenreHorror, GenreMystery, Runtime, RatedG, RatedM, RatedM/PG, RatedNC-17, RatedNot, RatedPG, RatedPG-13, RatedR, RatedTV-MA, RatedUnrated, Metascore, IMDB.Votes (Appendix A5). Furthermore, Figure 3 depicts lasso results, which completely agreed only with the Forward selection in terms of BIC, both indicating the best subset to have only 5 features: Year, GenreComedy, Runtime, MetaScore, and IMDb.Votes (Appendix A4).

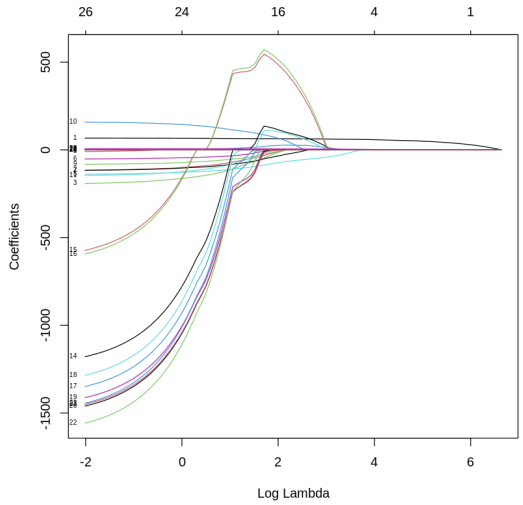


Figure 3. Results of Lasso Feature Selection indicating each of the coefficient paths as a function of the logarithmic of lambda values.

These results were relatively inconclusive as there was a discrepancy between the feature selection methods as well as across the criteria. Thus, we were left to make our decision of which regression set is optimal based on balancing utilizing a simpler model with fewer predictors that potentially would not capture all the variance of the outcome variable or using a much more extensive number of variables and risk overfitting the data.

We found that the only four variables constant across the entire analysis were Year, Runtime, Metascore, and IMDb.Votes. The additional variables in each of the larger regressions were different choices of the MPAA rating and the genre of the film. We thus decided to begin with the simplest model, including these four core variables as our baseline model and iteratively including the genre and rating one-hot encoded variables that were present in the backward and best subset selection models to understand which could improve the predictive power of the model.

B. Regression Analysis

To understand the relationship between the stock market and box office movie revenues, multiple regression analyses were performed to identify if there is a pattern between the economy and the film industry. A simple linear regression was used to establish if there is a linear relationship between stocks and movie profits. A polynomial model was then constructed to determine whether different stock closes over the years could be used to predict box office revenues. In addition to this, an SVM and a Logistic regression model were developed to assess if box office revenues could be used to predict stock market trends.

a. Modeling Film to Finance

The initial regression analysis conducted was a simple linear regression, to

explore whether stocks and box office revenues depicted a linear relationship; we explored yearly averages of stocks (focusing on the adjusted close) and the total yearly returns of box office revenue between the years 1999 - 2019. These values were aggregated to compare year-to-year performance and eliminate potential outliers from the initial raw day-to-day data. This preliminary analysis was chosen to determine whether a linear model could predict box office performance. A scatter plot and the results of the fitted model are represented in the results section with the R^2, which represents how the stock market explains the variance in the box office data.

Upon establishing the relationship outlined above; a polynomial regression was chosen to capture potential non-linear patterns within our data. This regression model was chosen because it allows greater flexibility than traditional linear regression, and can model more complex data trends. It is particularly strong when fitting on variables that are curved but still experience smooth & aligned patterns. The initial baseline model included all of the stocks in our dataset as predictors of box office revenues. The secondary model incorporated selective stocks identified in the feature selection process as the strongest predictors and was created with the goal of enhancing model performance and reducing noise.

- Baseline Model: 'Year', 'Adj Close' of all 32 stocks as predictors for 'Box Office'
- Selective Stock Model: 'Year', 'Adj Close' – SPX, XOM, T, GOOG, and LDOS as predictors for 'Box Office'

Prior to fitting either model, the dataset was split into a 70/30 training and test split. In addition, we focused on lower degrees for the models: 2 & 3. During the testing of these models, higher-degree polynomials performed poorly and failed to capture the trends within the test set. The results of the baseline and refined models are showcased in Figures 6 and 7.

b. Modeling Finance to Film:

We rigorously tested several regression models, including Support Vector Machine (SVM), logistic regression, linear regression, and XGBoost, to comprehensively understand the relationship between the stock market and the film industry. This approach allowed us to evaluate how well movie features could predict stock market movements. Recognizing that stock prices are not independent daily, we created lagged stock features in the dataset to capture historical momentum and trends. Specifically, we added the previous day's closing price, the closing price from three days prior, the percentage change from the previous day, the three-day percentage return, and the five-day percentage return to provide a broader window into stock price behavior following a film's release. To better model the movie features, we applied a log transformation to the IMDB Votes variable to compress its skewed distribution, making it more suitable for analysis. We also created dummy variables for categorical features, encoding the film's rating and primary genre into separate binary columns. Additionally, we ensured that all key dummy variables were present across the dataset, even if some categories were missing from the sample, to maintain consistency in the feature set.

We applied an 80/20 train-test split for all models, standardized the features using standard scaling, and performed five-fold cross-validation. We started with a baseline of four movie features and iteratively expanded to broader feature sets across three model types:

- Baseline Model: 'Year,' 'Runtime,' 'Metascore,' 'IMDB Votes'
- Most Inclusive Model: 'Year," 'GenreComedy," 'GenreDrama," 'Runtime," 'RatedG," 'RatedM/PG," 'Metascore," 'IMDB Votes'
- Rating-Focused Model: 'Year," 'Runtime," 'RatedG," 'RatedPG," 'RatedPG-13', 'Rated-R," 'Metascore," 'IMDB Votes'

For the Support Vector Machine (SVM) model, we used an SVM with a polynomial kernel of degree two. The polynomial kernel was chosen to capture nonlinear relationships and feature interactions (e.g. year and runtime) while maintaining a balance between model complexity and flexibility. We also balanced class weights for the imbalance between "stock down" and "stock up" cases and applied hyperparameter tuning with a regularization parameter of $C = 0.5$.

For logistic regression, we tested both Lasso (L1) and Ridge (L2) regularization and ultimately selected Ridge regularization for better stability. We used the liblinear solver, which is efficient for smaller datasets and supports L2 penalties. Class weights were also balanced to address the class imbalance.

We modified the target variable for linear regression to reflect continuous stock price movement rather than binary direction. We tested L1 and L2 regularization, tuning the alpha parameter from 0.01 to 1.0, and selected an alpha of 1.0 to control model complexity.

Finally, we selected XGBoost for its ability to capture complex, nonlinear relationships, automatically model feature interactions, and handle outliers and imbalanced data effectively. Using GridSearchCV, we identified the best hyperparameters for XGBoost as: colsample_bytree: 1, learning_rate: 0.05, max_depth: 7, n_estimators: 100, subsample: 1.

Section V. Results

A. Finance Response

The initial comparison between the stock market and movie box office revenues is represented in Figure 4. It is shown that both sectors have performed well over the years and depicted an upward trend since 1999, however, the box office experiences a volatility that the stock market does not consistently mirror. The only parallel movement observed between the two sectors is a short positive trend between 2008 and 2013, where both experienced a steady increase in their returns. This constant rise in the sector's profits may have been credited to the post-recession economic recovery period. Beyond this slight correlation, both variables' trends seemed independent, which suggested that if a potential relationship existed, it may not be strongly linear. The result of the figure led us to conduct a linear regression analysis to test the strength and direction of the relationship.

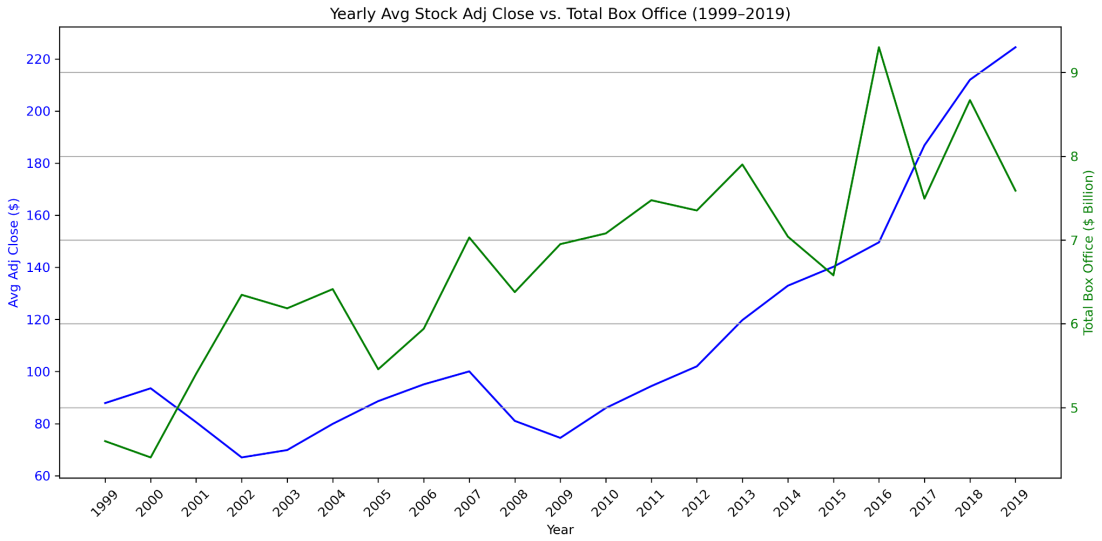


Figure 4. Yearly Average Stock Adjusted Close vs Total Box Office

A preliminary analysis was done using a simple linear regression to explore whether there is a linear relationship between average stocks and movie box office revenues – represented in Figure 5. A scatterplot was produced with a fitted regression line, which revealed a slight upward trend, suggesting a potential positive linearity between the two variables. The regression model produced an R^2 of 0.35, indicating that approximately 35% of the variance in box office data could be explained by the changes in the average adjusted stocks. Although the model indicated some correlation, it is relatively weak to conclude predictions, meaning that the remainder of the 65% of the variance could be influenced by other non-linear factors. The results of the initial linear regression did not support the presence of a strong linear relationship, which prompted us to explore a more flexible model like polynomial regression.

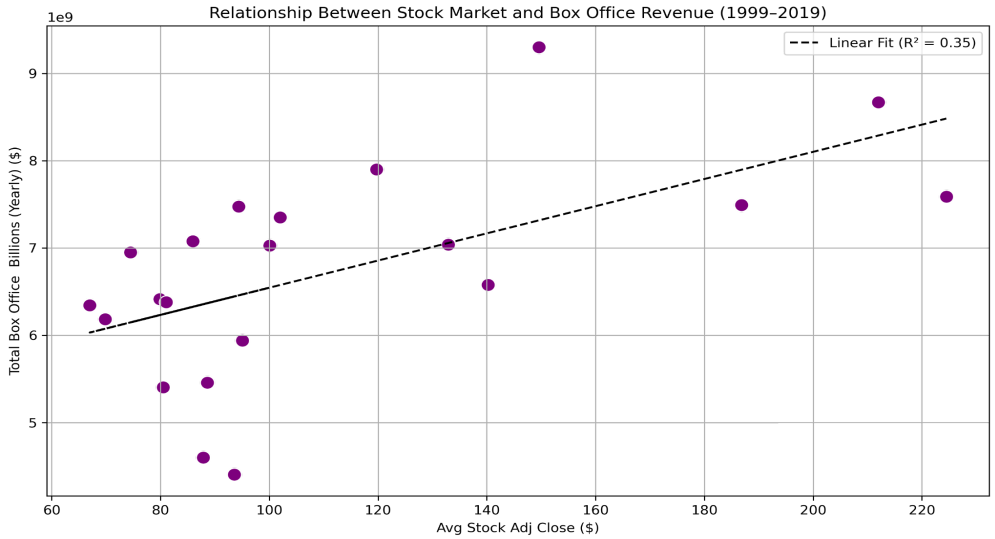


Figure 5. Relationship Between Stock Market and Box Office

Next, the data was modeled utilizing a polynomial regression to capture potential non-linear patterns. The initial baseline, represented in Figure 6**,** used a quadratic and cubic model (second-degree and third-degree polynomials). Lower degree polynomials were selected because higher degree models overfit to the data during testing. These models failed to generalize and performed poorly in predicting box office revenues. In the baseline model, the second-degree polynomial performed the best with an R^2 of 0.408, while the third-degree polynomial underperformed with an R^2 of -2.772.

This baseline model was then refined by reducing the number of stock predictors. We focused on the stocks identified in the feature selection process: SPX, XOM, T, GOOG, and LDOS. Figure 7 shows the updated scatter plot from Figure 6, with the refined model. This model consisted only of the second-degree polynomial since it had previously outperformed the third-degree polynomial. The refined model did not produce the expected improvement as its performance dropped to an R^2 of -0.71.

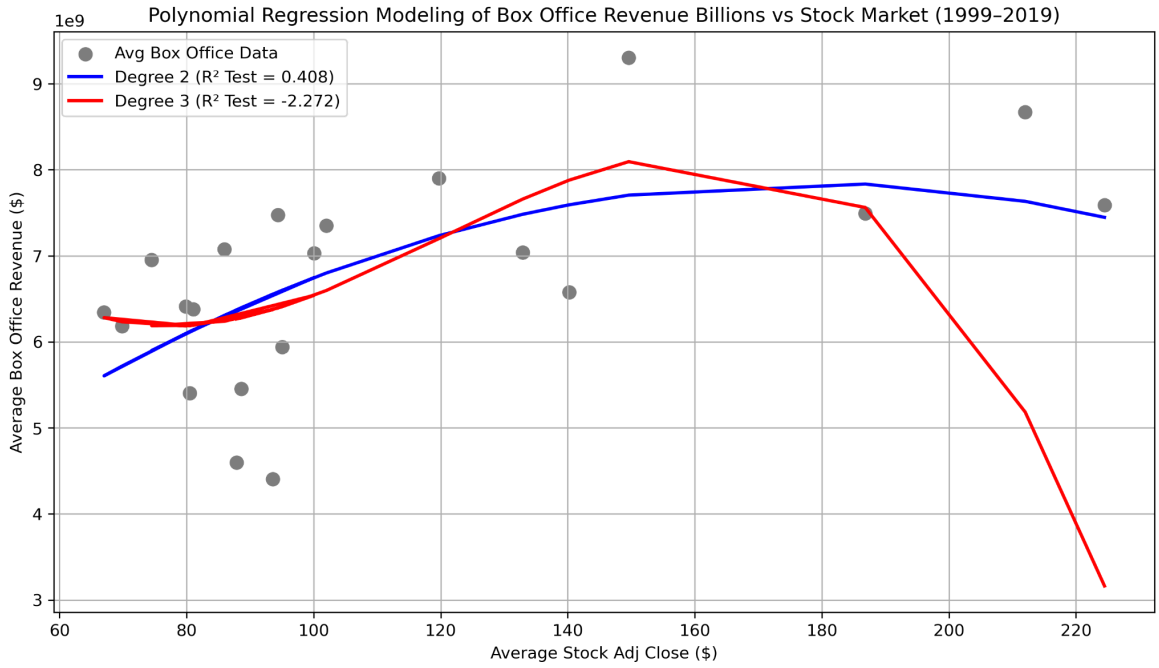


Figure 6. Baseline Polynomial Regression Modeling of Box Office vs Stock Market

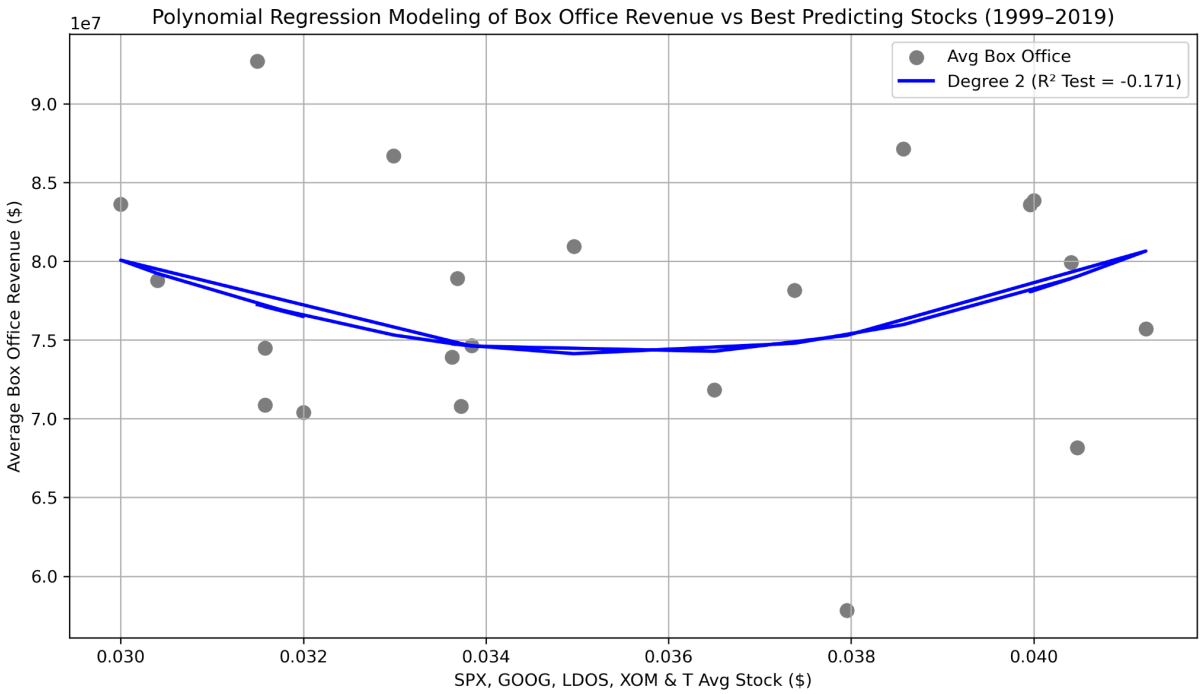


Figure 7. Refined Polynomial Regression Modeling of Box Office vs Stock Market

B. Film Response

In Figure 8, we present a Receiver Operating Characteristic (ROC) curve comparing logistic regression (green) and SVM with a polynomial kernel (blue) using the baseline feature set. The logistic regression model achieved an AUC score of 0.569, while the SVM model achieved a slightly higher AUC of 0.589. Since an AUC of 0.5 represents random guessing and 1.0 represents a perfect model, both models demonstrate only marginal improvement over random chance. Although the SVM slightly outperforms logistic regression, the overall predictive power remains weak. Adding more features to the models did not meaningfully improve the AUC scores, suggesting that predicting next-day stock movements using these movie and stock features is inherently complex. This result highlights the limited predictive power of the available features in a binary classification framework.

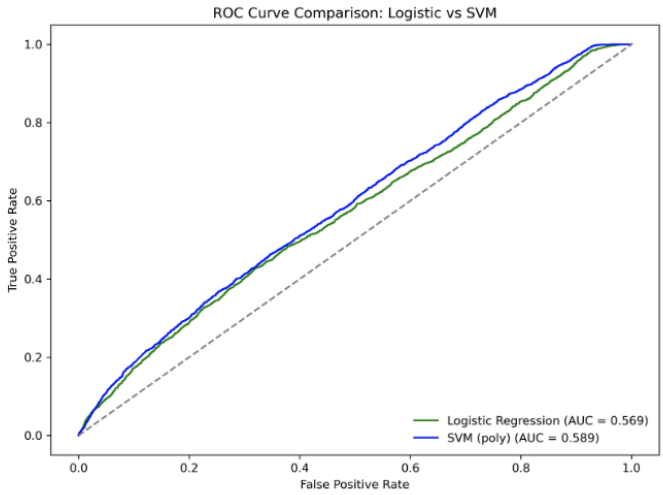


Figure 8. ROC curve comparison of Logistic regression and Support Vector Machine

Figure 9 is a scatter plot of the linear regression model predicting stock price movement using the baseline feature set. We focused on the baseline model because adding more features worsened the model's performance. Most predictions cluster around zero, even when the price change is substantial. When the actual price changes are extreme, the model tends to underpredict, pulling all predictions back toward the center. The model particularly struggles to capture large downward movements. Overall, the linear regression model is biased toward predicting small changes and fails to capture significant swings in stock behavior. This result is consistent with the model's R-squared value of 0.17, indicating that it explains only about 17% of the variance. The RMSE of approximately 290.0 suggests that, on average, predictions are about $290 off, while the MAE of 110.51 reflects a typical prediction error of around $110. These results suggest that the underlying data is not well-suited for a simple linear model. Stock price movements based on movie and stock features exhibit more complex, nonlinear patterns that linear regression cannot fully capture.

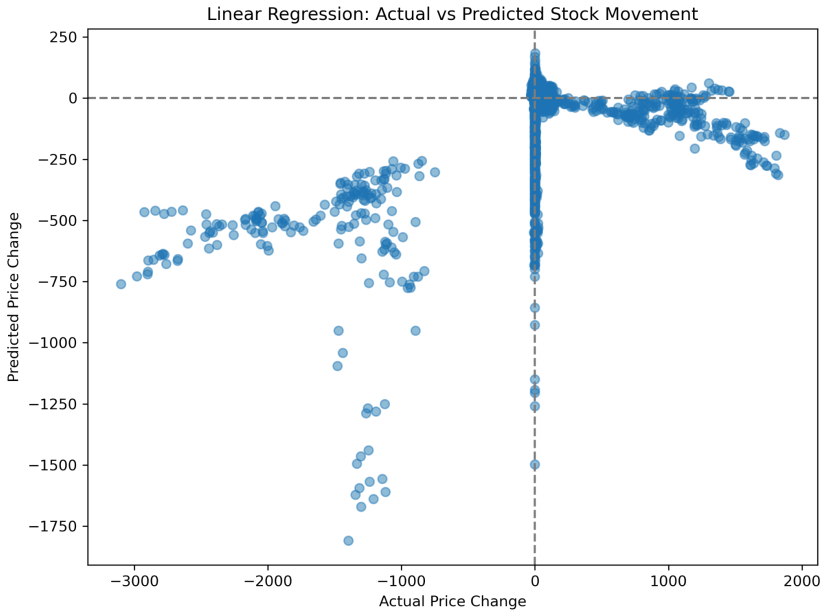


Figure 9. Linear Regression Scatter Plot of actual vs Predicted Stock Movement

The scatter plot of the tuned XGBoost model predicting stock price movements using the baseline feature set is depicted in Figure 10. We used the baseline features, as adding more variables again worsened the model's performance. From the evaluation results, we obtained an RMSE of 155.28, indicating that the typical prediction error is about $155. The MAE of 40.17 reflects a median prediction error of around $40, and the R-squared value of 0.76 means the model explains approximately 76% of the variance—a strong result. In the scatter plot, the points are tightly clustered along the diagonal line, showing that the model's predictions closely track the actual stock movements. The model handles large positive and negative swings effectively, with much less bias toward zero than the linear regression model. Although some noise remains near zero, it is significantly reduced. Examining the feature importance rankings from XGBoost (Appendix A6), the top three predictors are lagged stock features, led by the five-day return. Movie features, such as IMDB Votes and Metascore, contribute a more minor but meaningful boost. This demonstrates that stock-related features dominate prediction performance, while movie features add supplementary information. These results show that XGBoost captures the relationship between movie and stock features and stock price changes much more accurately than simpler models. The findings suggest that stock behavior based on movie-related and lagged stock features follows complex, nonlinear patterns that XGBoost can model effectively.

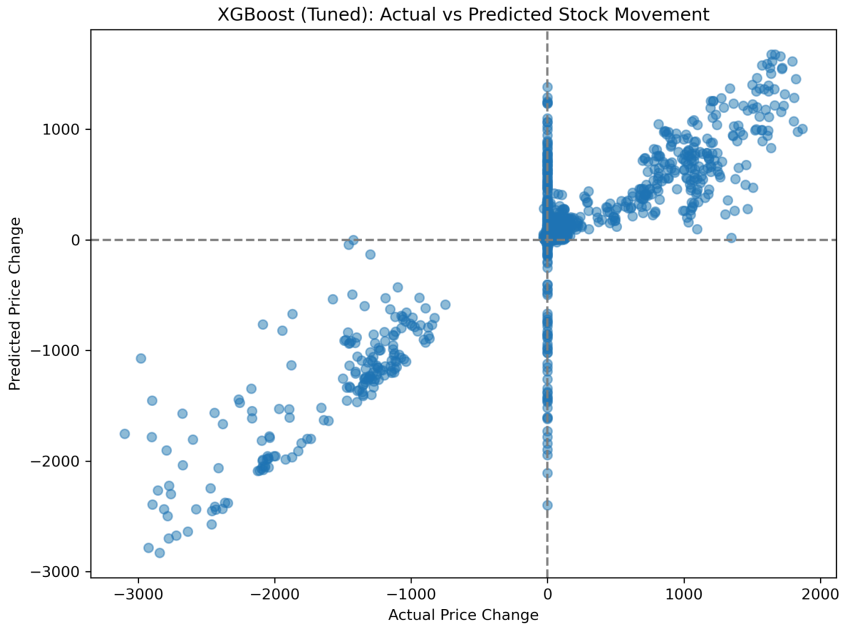


Figure 10. XGBoost scatter plot showing the Actual vs Predicted Stock Movement

In conclusion, we found that binary classifiers such as logistic regression and SVM performed poorly in predicting whether the stock market would go up or down based on movie features. Similarly, the linear regression model performed poorly, suggesting that the relationship between the features and stock movements may not be well captured by a linear model. Due to the complex and nonlinear nature of the dataset, it makes sense that the XGBoost model performed the best, successfully capturing some swings in stock movement based on movie and stock-related features. However, we could not develop a consistently strong predictive model even after splitting the data, applying cross-validation, performing feature selection, adding regularization, and hyperparameter tuning. Adding additional features either worsened or had no meaningful effect on performance, highlighting how difficult it is to predict stock movements, and ultimately suggesting that movie data alone is insufficient for reliably forecasting the stock market.

Section VI. Discussion

The tested models proved to consistently underperform, making them difficult to extrapolate for further analysis. Future work should focus on refining the connection between the film industry and the stock market by improving the model structures, examining the correct predictors, and accounting for time dependencies. For example, Recurrent Neural Networks might be able to capture the time dependencies more effectively than our models did. Although the dates were not included in our models, adding them as a feature in an LSTM or GRU may help the model learn patterns over time instead of simply learning patterns across films. We

eliminated dates as a feature to minimize dependency between observations, but it is difficult to eliminate the time feature from a time series dataset.

To further emphasize this point, we ran an Auto-Correlation Function on the SPX index and box office revenues. As we can see in Figures 11 and 12 below, there is a high level of autocorrelation in both variables due to the time series nature of our data. SPX was highly correlated across all twenty lag values. Twenty was chosen as an arbitrary hyperparameter, but we observed no significant changes across these lags so no further exploration was needed. However, with the box office revenues, certain lags limited autocorrelation. This plot showed us that incorporating moving averages across various films as a feature in our modeling would help eliminate this autocorrelation. Various Recurrent Neural Networks would improve our findings as they take this into account. An ARIMA model is also a logical next step to improve results.

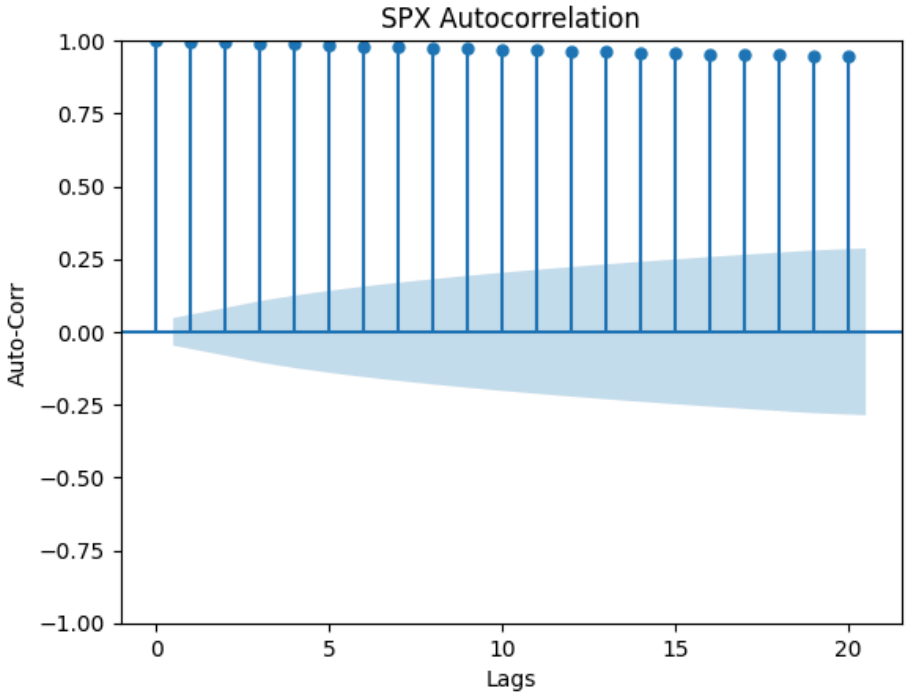


Figure 11. Autocorrelation between SPX observations across different lags

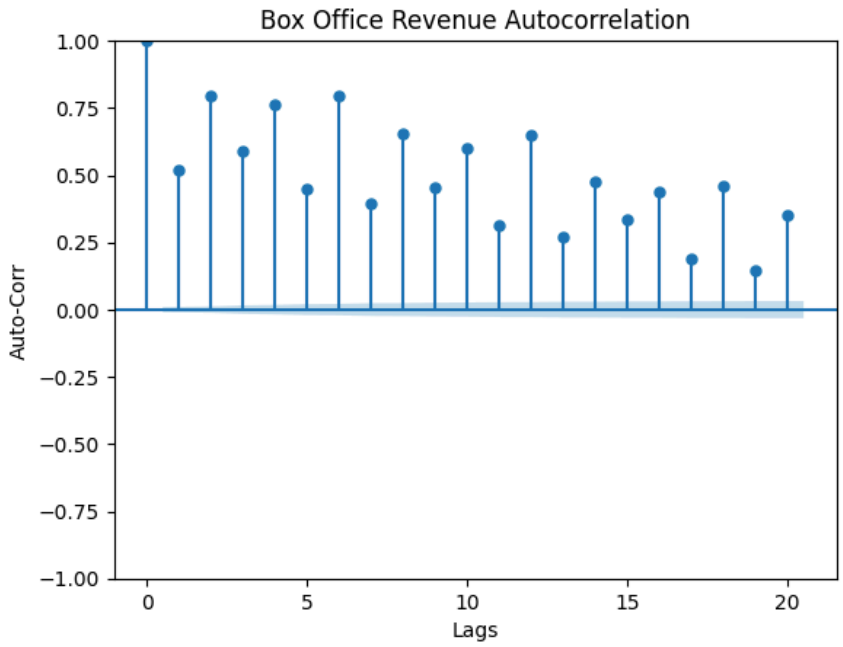


Figure 12. Autocorrelation between Box Office Revenue observations across different lags

Another limiting factor is simply the complexity and sometimes sporadic nature of the stock market and box office revenues. Both are affected by external factors, from current events to the rise of streaming platforms. The models will continue to struggle without context on the outside world that may be pushing revenue or prices up and down. However, this context is increasingly difficult to encode outside of MLP.

Increased feature engineering would help reduce noise and improve the model's performance. Adding technical indicators like Moving Average Convergence Divergence (MACD) and Relative Strength Index (RSI) could enhance the model's ability to detect trends in the stock market. The MACD measures the difference between short- and long-term moving averages, helping identify shifts in market momentum. When the MACD line crosses above the signal line, it signals upward momentum, and vice versa for downward momentum. The RSI gauges whether a stock is overbought (above 70) or oversold (below 30), highlighting potential reversal points. Including these indicators would allow the model to capture market trends more effectively and improve its predictions by accounting for both momentum and market sentiment.

In terms of films, using the Genre Popularity Index, a feature indicating the popularity of the cast, or employing NLP techniques on the film's description, would result in a similar improvement. In addition, engineering moving averages across box office data would eliminate much of the autocorrelation that is hindering our models.

In sum, our models underperformed, however, many avenues of research exist to improve upon the results with further iterations.

Section VII. Conclusion

Our analysis investigated the relationship between stock market trends and the movie box office revenues, and whether their relationship could be leveraged to predict the performance of both variables using regression and classification models. While our preliminary regression analysis exhibited promising results, the performance of the subsequent models significantly declined. Approaches ranging from linear regression with the addition of polynomial features to logistic regression and SVMs struggled to accurately predict stock market returns using box office revenues alone, and vice versa. Although models like XGBoost yielded improved R^2, other techniques that incorporated the inclusion of strong predictors, select stocks, and movie features identified in feature selection, like the second-degree polynomial, failed to generalize. The results obtained from our refined model prompted an additional Auto-Correlation analysis that revealed significant dependencies in both SPX and the box office revenues. This analysis suggested that other models such as ARIMA or Recurrent Neural Networks may work better to capture trends in our dataset than traditional regression. Ultimately, the behavior of both the stock and movie sectors appears to be influenced by independent factors, with low R^2 and higher RMSE & MAE values, our models' results challenged our hypothesis that there is a strong interconnectedness between the two industries.

References

Carollo, L. (2024, February). *Estimated revenue of the motion picture and video production and distribution industry in the United States from 2005 to 2022*. Statista. https://www.statista.com/statistics/184140/estimated-revenue-of-us-motion-picture-and-video-industry-since-2005/

Chhajer, P., Shah, M., & Kshirsagar, A. (2022). The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction. *Decision Analytics Journal, 2*, 100015. https://doi.org/10.1016/j.dajour.2021.100015

IMDb. (2025, April 2). *IMDb ratings FAQ*. IMDb.com. https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV?showReportContentLink=false&reportContentLinkPath=%2Fcontact%2Freport#

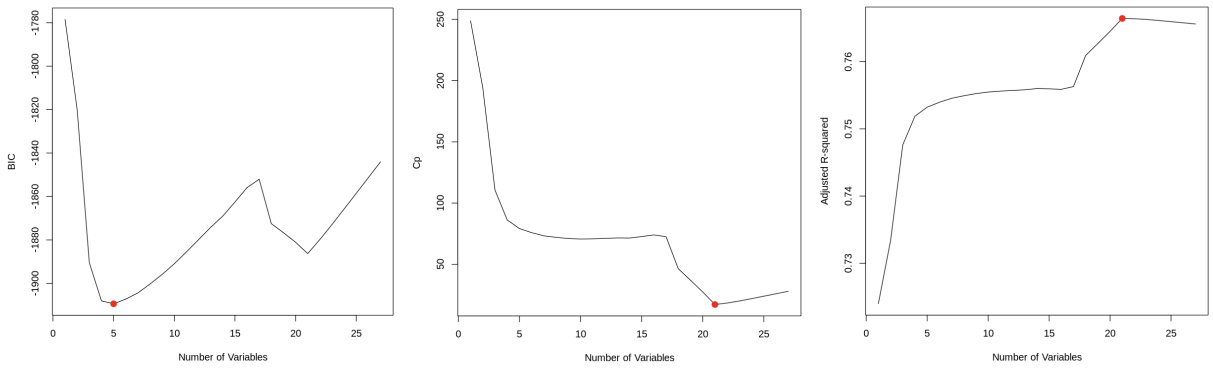
IMDb. (2014). *Movies with highest Metascore*. IMDb.com. https://www.imdb.com/list/ls051211184/

Einav, L., & Ravid, S. A. (2009). Stock market response to changes in movies' opening dates. *Journal of Cultural Economics, 33*(4), 311–319. https://doi.org/10.1007/s10824-009-9093-7

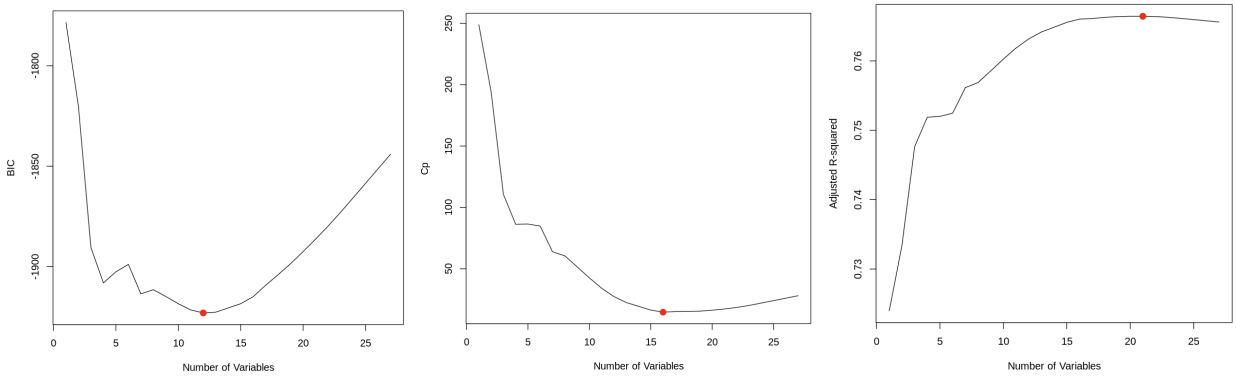
OMDb API. (n.d.). *The Open Movie Database*. Retrieved April 27, 2025, from https://www.omdbapi.com/

Statista. (2025, February). *Stocks - United States*. Statista. https://www.statista.com/outlook/fmo/stocks/united-states

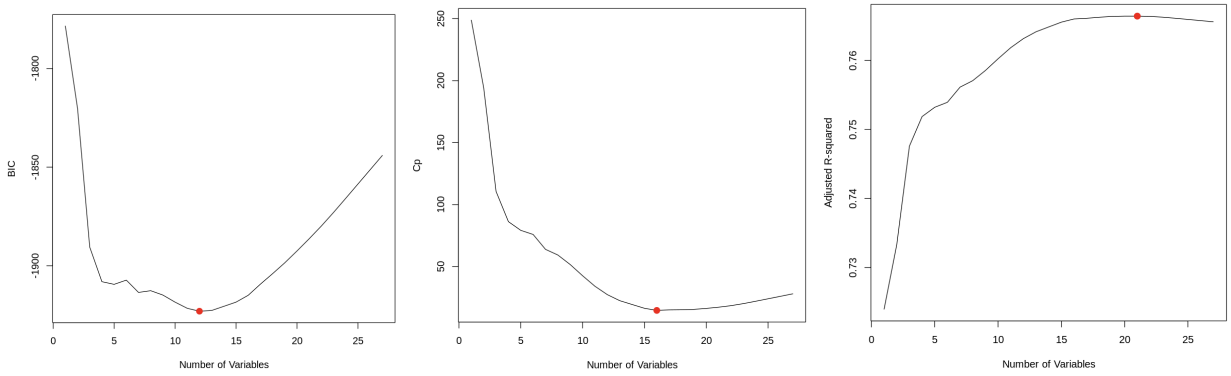
Appendix



A1. BIC, CP, Adjusted R results for Forward Stepwise Selection indicating the optimal number of variables with a red point.



A2. BIC, CP, Adjusted R results for Backward Stepwise Selection indicating the optimal number of variables with a red point.



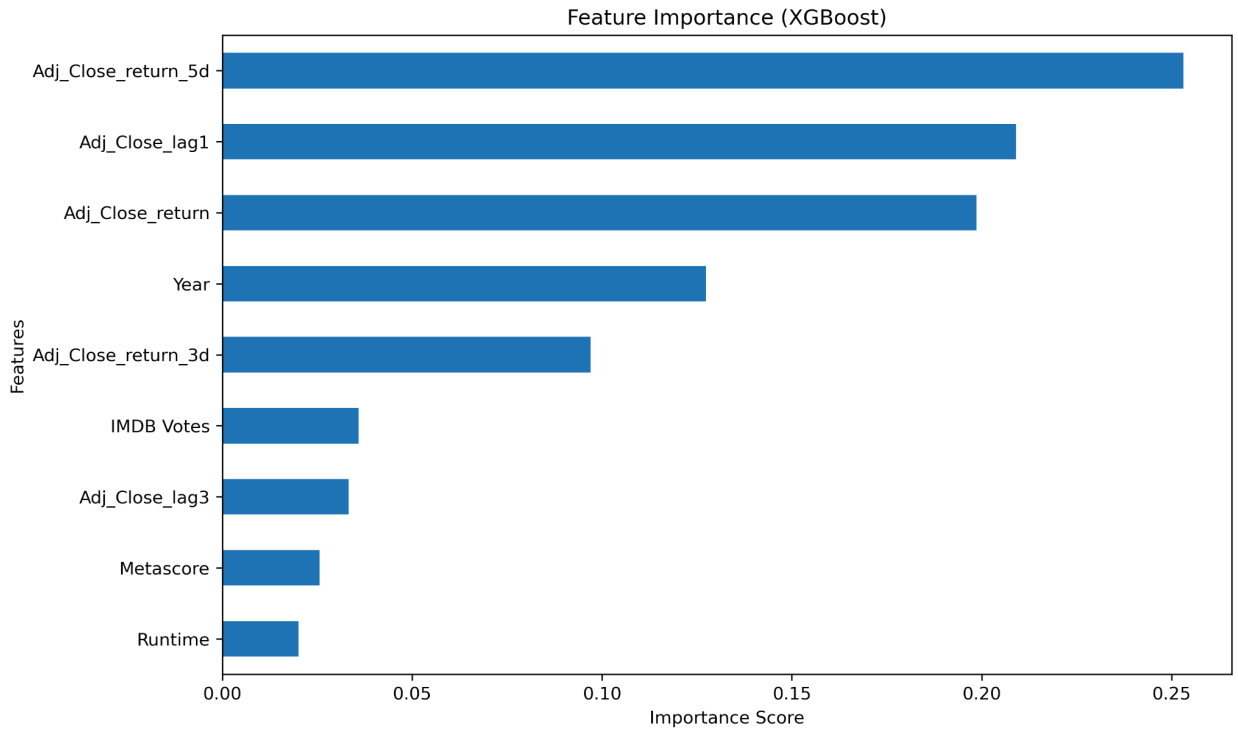
A3. BIC, CP, Adjusted R results for Best Subset Selection indicating the optimal number of variables with a red point.

	s1
(Intercept)	-1.180926e+05
Year	5.949273e+01
GenreAdventure	.
GenreAnimation	.
GenreBiography	.
GenreComedy	-3.084315e+00
GenreCrime	.
GenreDrama	.
GenreFantasy	.
GenreHorror	.
GenreMystery	.
GenreSci-Fi	.
GenreThriller	.
Runtime	1.048787e+00
RatedG	.
RatedM	.
RatedM/PG	.
RatedNC-17	.
RatedNot Rated	.
RatedPG	.
RatedPG-13	.
RatedR	.
RatedTV-MA	.
RatedUnrated	.
IMDB.Rating	.
Metascore	3.137034e+00
IMDB.Votes	-3.549603e-04
Box.Office	.

A4. Results for Lasso Selection showing the five selected variables along with the values for each of their coefficients.

	BIC	CP	Adjusted R^2
Best Subset	Year, GenreComedy, Runtime, MetaScore, IMDB.Votes	Year, GenreAdventure, GenreAnimation, GenreComedy, GenreCrime, GenreDrama, GenreHorror, GenreMystery, Runtime, RatedG, RatedM, RatedM/PG, RatedNC-17, RatedNot, RatedPG, RatedPG-13, RatedR, RatedTV-MA, RatedUnrated, Metascore, IMDb.Votes	Year, GenreAdventure, GenreAnimation, GenreComedy, GenreCrime, GenreDrama, GenreHorror, GenreMystery, Runtime, RatedG, RatedM, RatedM/PG, RatedNC-17, RatedNot, RatedPG, RatedPG-13, RatedR, RatedTV-MA, RatedUnrated, Metascore, IMDb.Votes
Forward Stepwise	Year, GenreComedy, Runtime, MetaScore, IMDB.Votes	Year, GenreAdventure, GenreAnimation, GenreComedy, GenreCrime, GenreDrama, GenreHorror, GenreMystery, Runtime, RatedG, RatedM, RatedM/PG, RatedNC-17, RatedNot, RatedPG, RatedPG-13, RatedR, RatedTV-MA, RatedUnrated, Metascore, IMDb.Votes	Year, GenreAdventure, GenreAnimation, GenreComedy, GenreCrime, GenreDrama, GenreHorror, GenreMystery, Runtime, RatedG, RatedM, RatedM/PG, RatedNC-17, RatedNot, RatedPG, RatedPG-13, RatedR, RatedTV-MA, RatedUnrated, Metascore, IMDb.Votes
Backward Stepwise	Year, Runtime, RatedG, RatedNC-17, RatedNot, RatedPG, RatedPG-13, RatedR, RatedTV-MA, RatedUnrated, Metascore, IMDb.Votes	Year, GenreAdventure, GenreAnimation, GenreComedy, GenreDrama, Runtime, RatedG, RatedNC-17, RatedNot, RatedPG, RatedPG-13, RatedR, RatedTV-MA, RatedUnrated, Metascore, IMDb.Votes	Year, GenreAdventure, GenreAnimation, GenreComedy, GenreCrime, GenreDrama, GenreHorror, GenreMystery, Runtime, RatedG, RatedM, RatedM/PG, RatedNC-17, RatedNot, RatedPG, RatedPG-13, RatedR, RatedTV-MA, RatedUnrated, Metascore, IMDb.Votes

A5. The optimal regression variables for each feature selection method and each criterion.



A6. Feature Importance from XGBoost Model