Predicting Box Office Success: Enhancing Accuracy with Keyword Analysis

Isabel O'Connor, Katie Shaughnessy, Emmanuella Cann, Carson Colyer, Krishu Wadhwa

Final Paper

Professor Johnson

DS 3001-002

14 December 2024

Abstract..... **2**
Introduction..... **3**
Data..... **6**
Methods..... **13**
Results..... **17**
Conclusion..... **20**
References..... **22**

Abstract

For our research project, we sought to predict box office success using indicators such as director, production company, genre, and budget. Our initial model was produced without the incorporation of keywords, using the explanatory variables “director”, “genre”, and “season” (of movie release). The model was created using a 70/30 train/test split, yielding a 0.586 R^2 value for the test set. A similar R^2 value of 0.545 for the training set suggests that the model was not prone to overfitting or underfitting. However, the R^2 values also suggest that a mere half of variation in box office success is explained by our chosen variables.

In the second stage of analysis, we added keywords to the model in an attempt to improve predictive accuracy. Subsequently, we selected seven keywords from a list of 40 movie keywords associated with the highest average box office revenues. A boolean variable “keywords_encoded” was then added to the original model, taking a value of 1 if any of the selected keywords were present in a movie’s description, or 0 if none were present. We observed that compared to the previous model, the Test R^2 improved to a value of 0.634, while the training R^2 decreased to 0.510. This indicates underfitting in the model.

Overall, we found that the addition of keywords (with the method used) had an ambiguous effect on overall model performance. We found that keywords associated with high average box office revenues were also generally associated with unique, high grossing movies (aka: “iceberg”, “titanic”, “transformers”, etc). Since these words are highly specific, they are not very useful for predicting general box office outcomes. To improve the model’s rate of predictive accuracy, we could incorporate a more extensive set of keywords, as well as engineer our model to better handle outliers in the dataset.

Introduction

Predicting box office success in terms of revenue is a complex challenge because it requires analysis of numerous features including genre, budget, season of release, studio, and keywords. Keywords, especially as textual data, have the potential to capture thematic, marketing, and audience engagement related elements that traditional numerical variables can not. However, as multi-faceted and high dimensional data, the predictive value of keywords remains unclear.

Our group was motivated to explore economic trends in the film industry– specifically, by identifying film characteristics which correlate to high box office success, and evaluating the predictive value of keywords. For this purpose, a dataset entitled “Movies Dataset: Budgets, Genres, and Industry Insights” was sourced from Kaggle, providing data for nearly 5,000 films. We aimed to leverage this dataset to extract meaningful insights about the factors that drive financial success in the film industry. This data is particularly interesting and useful to analyze as it allows us to quantify the impact of creative and financial decisions on a film’s success, with the goal of providing actionable insights for both filmmakers and industry analysts.

In order to test the correlation between movie characteristics and box office success, we developed plans for two comparative models. Model one would use a simple 70/30 training/test split in order to predict the impact of key variables (budget, genre, release season, and director) on box office success. Subsequently, we would identify the top 40 keywords associated with high average box office success. Several of these words would be selected, and model 2 would include a boolean variable equal to “1” if any of these words were present in a film’s description. This would allow us to test whether box office outcomes could be more accurately predicted with the addition of keywords. All other parameters of model 2 would remain constant from the original.

To compare the predictive accuracy of the two models, we planned to analyze metrics such as R^2 and RMSE values.

During exploratory analysis, we encountered nuances within our data which ultimately motivated our choice of feature design for model 1. For instance, we discovered that our dataset included thousands of movie directors, but the vast majority of these individuals had directed only one movie. Subsequently, we observed that the average box office earnings per director were extremely left skewed, indicating that only a handful of directors had completed high-grossing projects. A similar trend was observed with production companies, suggesting a relatively oligopolistic market structure within the film industry. To account for the skewed distribution of average revenue per director, we designed a "directors_encoded" variable, which took a value between 1 and 4 depending on the director's percentile placement for average box office outcome. In addition, the "genre" and "season" variables were engineered using one-hot encoding. During initial EDA, we found that movies released in the summer or winter seasons typically generate higher box office earnings. Thus, we encoded specific variables for "summer" and "winter" release, as well as an "other" variable for the off-season. The "genre" variable was then encoded using a dummy variable. In addition, a numerical "budget" variable was included.

After running model 1 using a 70/30 training/test split, we received a training R^2 value of 0.542, as well as a test R^2 value of 0.590. Furthermore, the training RMSE reported was 108,254,400.01, whereas the test RMSE was reportedly equal to 108,086,028.67. The model performed similarly between training and test data, with no major signs of overfitting or underfitting. The RMSE values suggest that, on average, the predictions for revenue are off by about 108.25 million for the training set and 108.08 million for the test set. The R^2 values

indicate that the model explains approximately 54.5% of the variance in revenue for the training data and 58.6% for the test data, showing a reasonable but not highly predictive model.

To account for high-dimensional data, we switched to the LASSO method when building model 2. For this iteration, we copied model 1's feature engineering for the variables "genre", "season", and "director". In addition to incorporating the "budget" numerical variable, we also added a boolean variable "keywords_encoded", which was designed to equal "1" if a movie contained any of the seven keywords 'universe', 'strength', 'floating', 'artifact', 'reef', 'shield', or 'fleet'. These words were chosen from a list of 40 movie keywords associated with the highest average box office outcomes.

When comparing model 2 to model 1, we found that the Training RMSE increased from 108,254,400 to 114,304,059 and Training R^2 decreased from 0.545 to 0.510, indicating that model 2 fit the training data slightly worse. However, the Test RMSE for model 2 was a significant improvement, dropping from 108,086,029 to 97,604,429. Furthermore, compared to model 1, the Test R^2 increased from 0.586 to 0.634, showing better generalization to unseen data.

The discrepancy between the training and testing metrics suggests that the keyword enhanced data has better generalization capabilities compared to the baseline model. While it may struggle to fully explain the variance in the training data, it avoids overfitting and leverages the added features to improve predictions in unseen data. These results reveal the potential value of keywords in predicting

Generally, the keywords chosen did little to increase the predictive strength of the model. To improve this outcome, we could include a greater variety of keywords in our model. Alternatively,

we could adopt a more robust strategy for selecting keywords – choosing only those which are simultaneously broadly applicable and associated with high-grossing films.

Data

Selection

We used a dataset titled *movie_dataset.csv*, which was sourced from Kaggle. This dataset contains information about 4,803 movies spanning various genres, production years, and countries of origin. The dataset originally includes 24 columns, each providing distinct attributes or metrics for the films.

Our primary research interest was predicting trends in movie success using indicators such as runtime, director, production company, and keywords associated with each film. The dataset was highly relevant as it included diverse information for each movie, eliminating the need for extensive data integration from multiple sources.

Key variables of interest included movie runtime, director, rating, genre, popularity score, and keywords. Specifically, we aimed to analyze whether certain keywords correlate with higher rates of popularity, user rating, or box office success. For instance, we hypothesized that films with keywords such as "Marvel," "fantasy," "comedy," or "adventure" were more likely to achieve significant box office success. As our dataset includes various movie-related indicators, we believed it to be well-suited for our research.

Predicted Challenges

One of the primary challenges in our research involved cleaning and managing outliers. For example, during our EDA, we identified several films with a runtime close to zero minutes and observed significant variability in runtimes overall. As a result, we needed to classify extremely short films separately from full-length features to maintain the integrity of our analysis.

Additionally, biases in certain variables posed challenges. For instance, the "popularity score" or "vote_average" variables could suffer from recency bias or discrepancies due to varying levels of internet accessibility at the time of a film's release. Addressing such biases was critical for obtaining accurate insights.

Key Variables

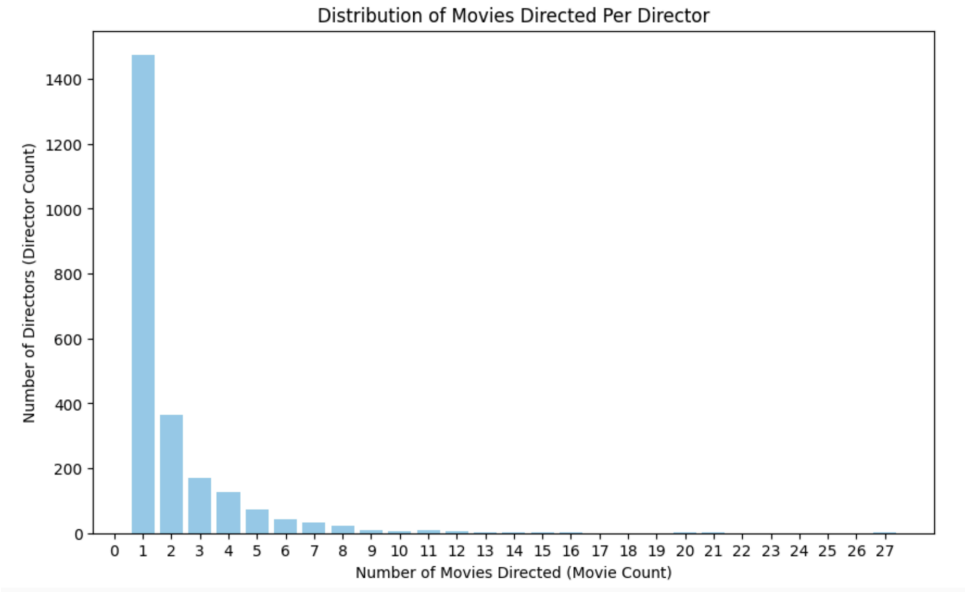
Variable Name	Description	Type
title	Title of the film	Object
budget	The movie's budget	Integer
genres	Genres of the movie	Object
keywords	Keywords associated with the movie	Object
popularity	Popularity score of the movie	Float
production_companies	Production companies involved	Object
production_countries	Countries where the movie was produced	Object
release_date	Release date of the movie	Object
revenue	Revenue generated by the movie	Integer

runtime	Duration of the movie (minutes)	Float
vote_average	Average user rating	Float
director	Director of the movie	Object

EDA Findings

Directors

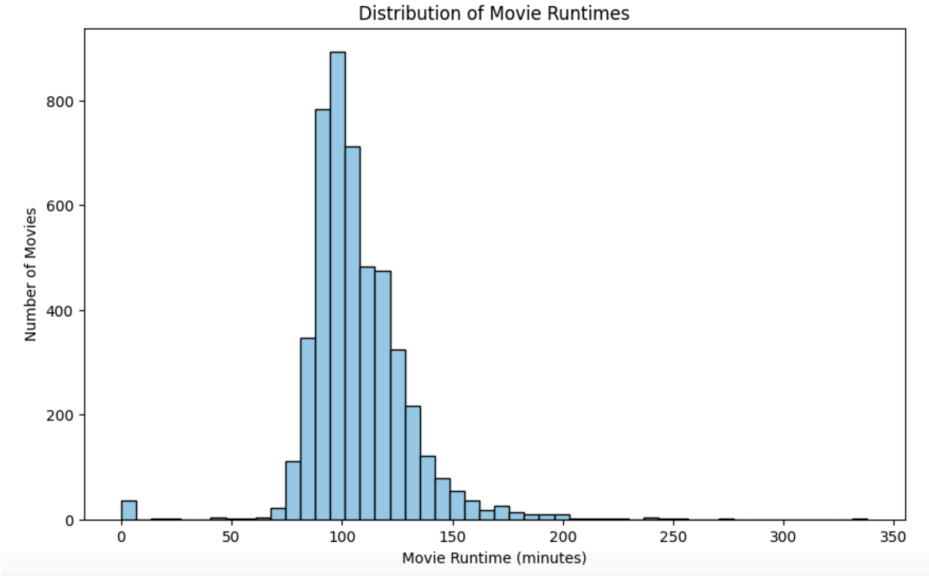
Analysis of the director variable revealed that most directors had only directed one movie in the dataset. The distribution of movies per director was strongly right-skewed, with a mean of 1.0 and a third quartile of 2.0. Outliers included Steven Spielberg (27 movies) and Woody Allen (21 movies).

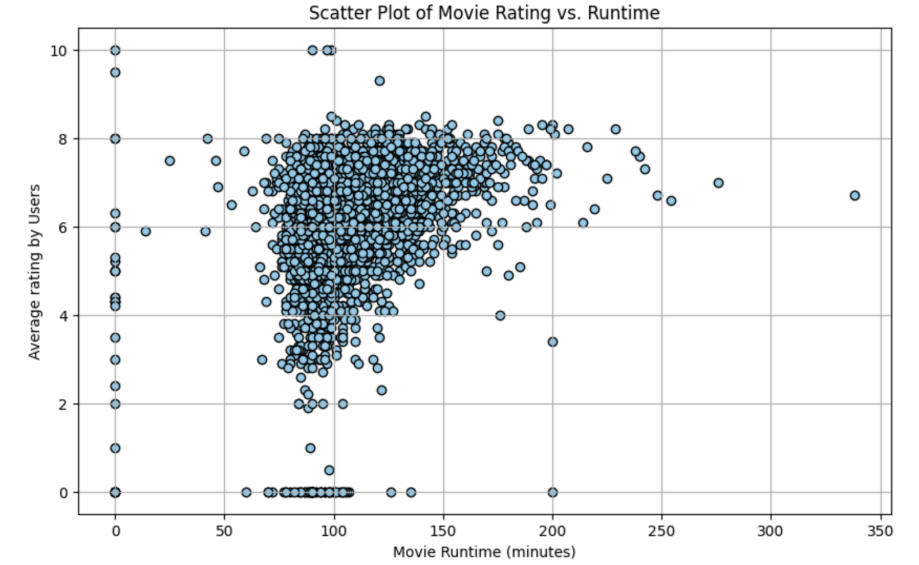


Movie Runtime

Movie runtimes varied significantly, with a unimodal distribution peaking around 100 minutes.

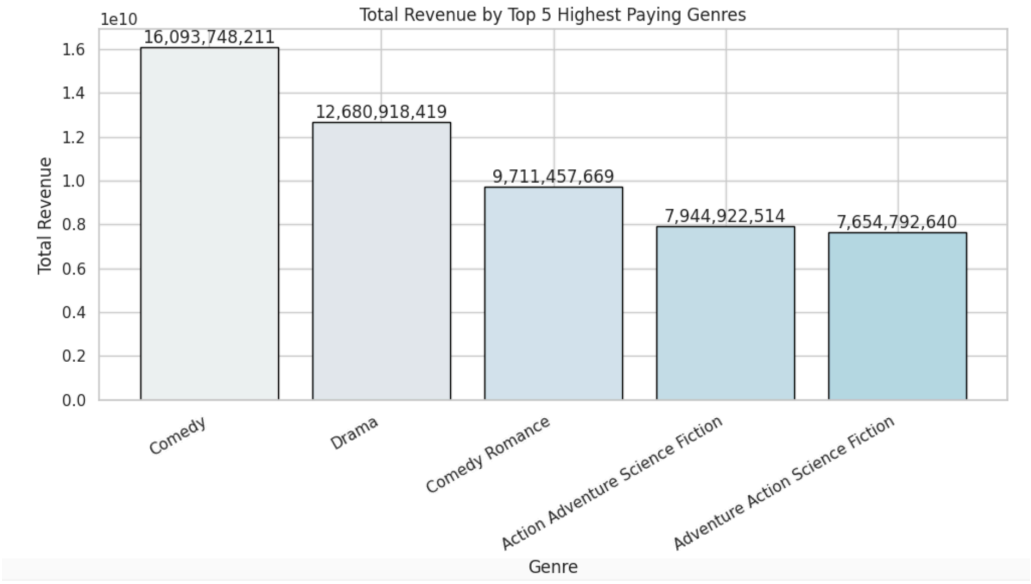
The median runtime was 103 minutes, with a standard deviation of 26 minutes and a maximum runtime of 338 minutes (over 5.5 hours). A scatter plot suggested a partially linear relationship between runtime and rating, with a correlation of 0.38.





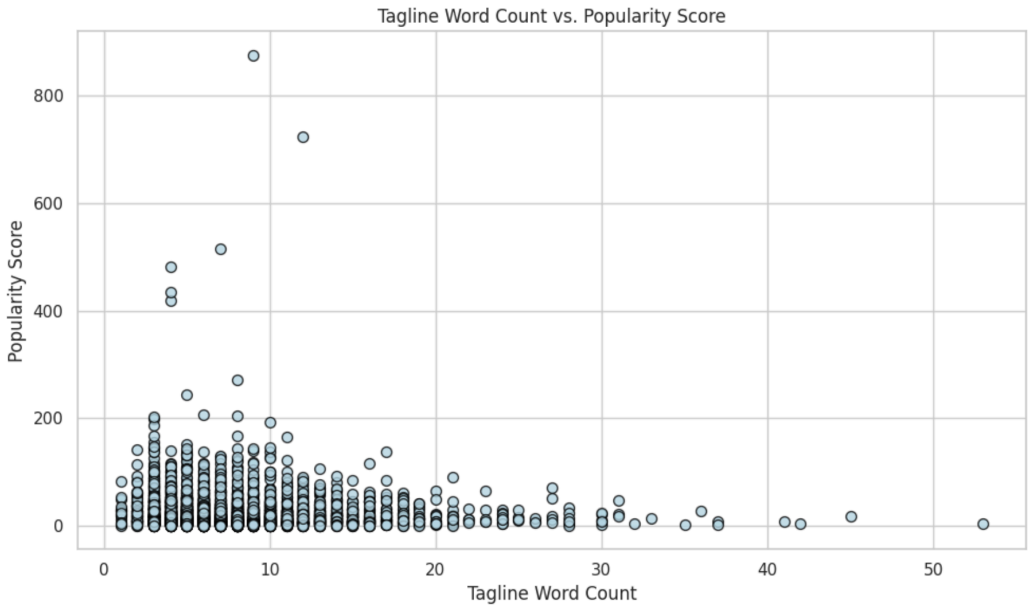
Genre and Revenue

Aggregate revenue by genre indicated that comedy had the highest total revenue, followed by drama, romantic comedy, and action/sci-fi films.



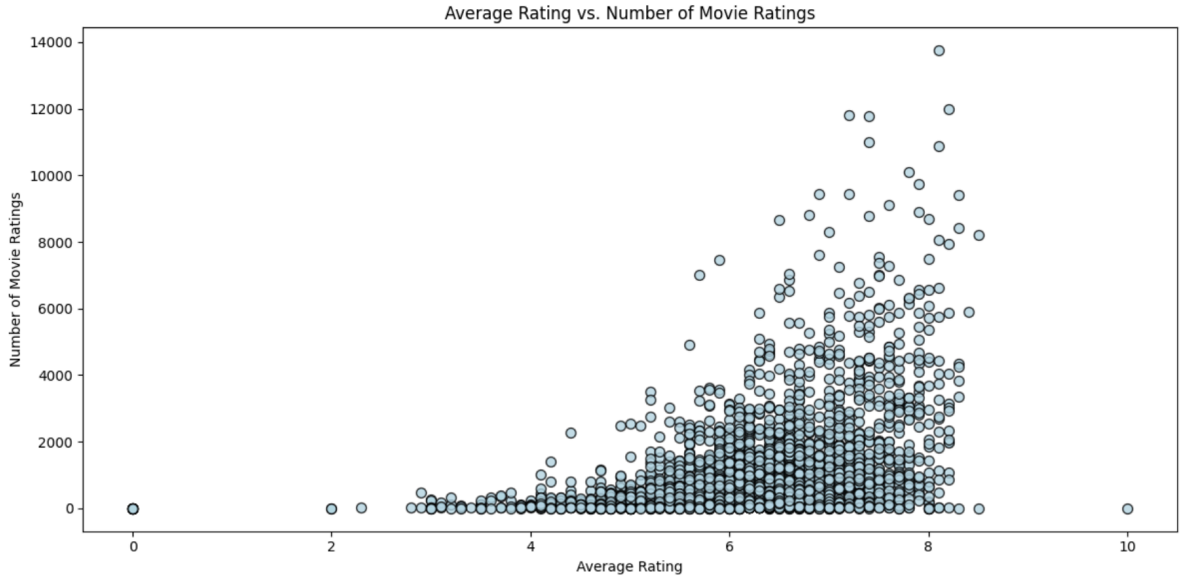
Tagline vs. Popularity

A graph comparing tagline length to popularity revealed a slight negative correlation, suggesting that shorter, "snappier" taglines may be more attractive to viewers.



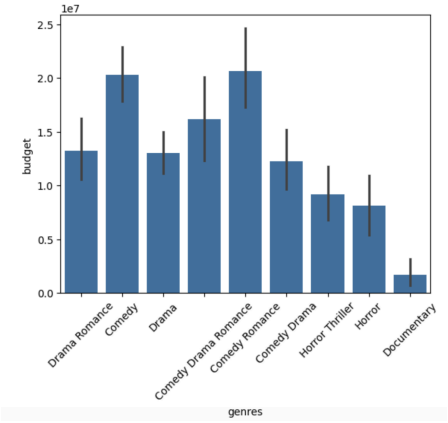
Average Rating by Number of Reviews

A scatter plot showed a positive correlation between average movie rating and the number of user reviews, suggesting that well-reviewed movies tend to attract more viewers and additional ratings.



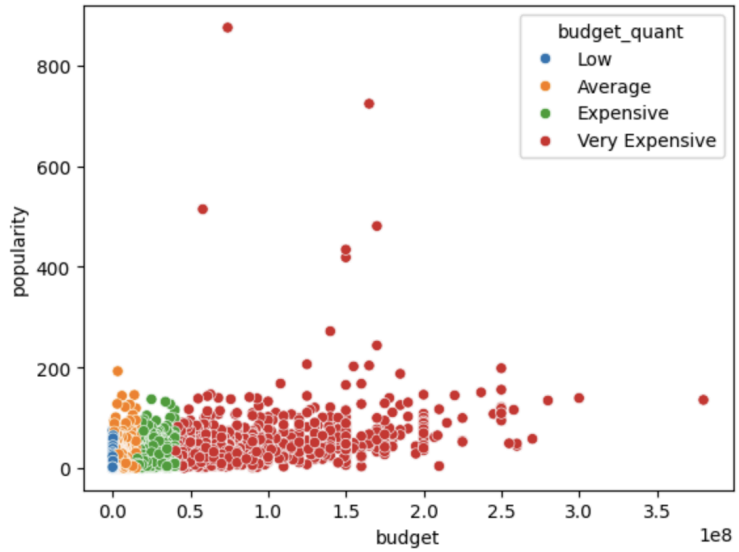
Budget vs. Genres

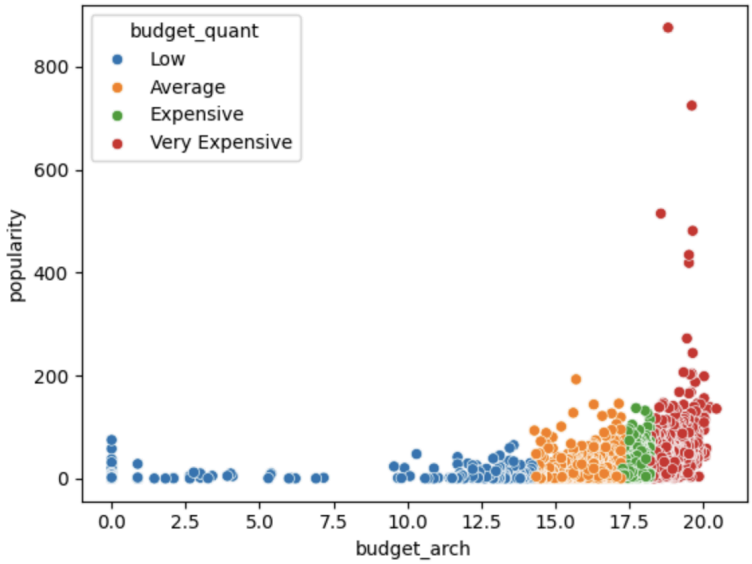
A bar plot revealed that among the ten most frequent genres, comedies and related genres tended to have the largest budgets, while documentaries had the lowest.



Budget and Popularity

A scatterplot comparing budget and popularity revealed a moderate positive correlation ($r = 0.50$), indicating that higher budgets often correlate with greater popularity. However, some outliers deviate from this trend.





Methods

Project Goal

For our project, we chose to create a predictive model for box office revenue (in thousands of dollars) using variables such as studio, genre, and other relevant factors. We then introduced keywords to assess whether they enhanced the model's predictive strength. We sought to analyze whether there are specific keywords associated with films which correlate to higher rates of popularity, user rating, and box office success. For instance, we predicted that movies associated with keywords such as "Marvel", "fantasy", "comedy", or "adventure" might enjoy greater box office success. Using supervised learning with a regression model, where each observation represents a single movie record, we aimed to capture the influence of these variables on box office outcomes.

To summarize, we sought to pursue answers for the following questions:

1. To what extent do keywords enhance the predictive accuracy (in terms of measured RMSE) of a box office success model built on traditional variables such as genre, studio, and budget?
2. Do genre-specific keywords add unique predictive value to a model for box office success, or are there universal keywords that enhance predictions across multiple genres?
3. How does the relative importance of keywords in predicting box office success vary based on the studio producing the film?

Strategy

A central question driving our analysis was the extent to which keywords improve the predictive accuracy of box office success models that traditionally rely on genre, studio, and budget.

Specifically, we aimed to measure improvements in model performance using metrics like R-Squared and RMSE (Root Mean Squared Error) to quantify the impact of keywords. Our investigation sought to explore whether genre-specific keywords add unique predictive value compared to universal keywords that may enhance predictions across multiple genres. Moreover, our analysis considered potential shifts in keyword relevance over time. Specifically, assessing how changes in audience preferences and industry trends impact the effectiveness of certain keywords. Finally, we aimed to understand how the importance of keywords in predicting box office success varies based on the producing studio, potentially offering insights into tailored marketing and production strategies.

Feature Engineering

After handling missing data through imputation or deletion, we split the dataset into training and testing sets. Prior to this, categorical variables like title, genres, keywords, production companies and production countries were one-hot encoded to provide insight into their relationship with the predicted variables. Initially, Principal Component Analysis (PCA) was to be used for numeric variables that are highly correlated. Furthermore, we planned to account for interaction effects, creating a single variable for highly correlated categorical variables. We also planned to perform any necessary transformations to our data to refine our model. To address potential overfitting from high-dimensional encoding, we applied regularization techniques like LASSO. Through these analyses and our planned feature engineering, we aimed to build a robust linear regression model that can effectively predict box office success based on the strategic use of keywords alongside other relevant variables. Through this project, we hoped to provide insights into movie trends and to inform effective marketing and production strategies based on data-driven analysis.

Measuring Results

To evaluate the model's accuracy, we used metrics like R-Squared and RMSE. The R-Squared value, or coefficient of determination, indicates the proportion of variance in box office revenue explained by our predictor variables, showing how well the model captures trends in the data. RMSE, on the other hand, measures the average magnitude of prediction errors, giving us insight into how close our model's predictions are to actual values.

Together, the R-Squared and RMSE metrics were used used to test the success of our model. Since models can be overfitted or underfitted, comparing the training and testing data RMSE can provide insight into how well our model will perform on novel data, as well as insights on how well our model is performing with current data. A high RMSE in the training and testing data indicates that our data is underfitted. On the contrary, a low RMSE in the training data and a low RMSE in the testing data indicates that the data is overfitted. A RMSE in which both the training and testing data RMSE are similar indicates a well-trained model. Additionally, the R-Squared provides insight into how much of the variance in our predicted value is explained by our predictors. A low R-Squared where there are multiple variables in our model signals us to a simpler, more refined model with more relevant variables.

Potential Design Weaknesses

Given the large number of variables in the movie dataset, it could prove difficult to design the optimal regression for predicting box office revenue. Some variables, such as budget and production company, should be interacted within the regression to ensure maximal predictive outcomes. To inform the variables, interaction terms, and power terms included in the regression model, we conducted a short meta-analysis to better understand the dynamics of movie revenue outcomes. If the model design still yields suboptimal predictive accuracy, further research and model engineering will be pursued.

Another potential challenge is our reliance on linear regression, which may not effectively capture the nonlinear and complex relationships affecting movie success. While factors like budget, director, keywords, runtime, and revenue play a role, there are also many other influences, such as cultural trends, social issues, marketing strategies, franchise popularity, and availability on streaming platforms or in theaters. To tackle this challenge, we will be using feature engineering to develop new variables that better reflect these complexities. For instance, we can create combined variables that look at the relationship between marketing spending and release timing to see how they impact box office results. Additionally, we can apply transformations to key variables, such as using logarithmic scales for budget and revenue, to help smooth out any extreme values. This way, we can improve our model's ability to understand the factors that contribute to a movie's success.

Additionally, certain categorical variables in the dataset have large numbers of potential values, which could affect the difficulty of dummy variable transformation. For instance, the “director” variable has a broad range of values, which could result in “messy” one-hot encoding. To avoid

this, we could engineer the director dummy variable such that directors associated with a large revenue impact (ex: Spielberg, Nolan, or Tarantino) are given their own value within the dummy variable, whereas all other directors are lumped into the “0” value. This method could also be applied to other categorical variables, where appropriate.

Results

Baseline Model

The baseline model, which used traditional variables such as genre, studio, budget, and season of release, achieved the following metrics:

- Training R^2: 0.545
- Test R^2: 0.586
- Test RMSE: $108.08 million

The similarity between training and test R^2 values indicates no major overfitting or underfitting. However, the RMSE of over $100 million highlights the challenge of accurately predicting box office revenue with these variables alone.

Keyword-Augmented Model

Adding seven keywords ("universe," "strength," "floating," etc.) as binary features improved the model's performance:

- Training R^2: 0.510 (decrease)
- Test R^2: 0.634 (increase)
- Test RMSE: $97.60 million (improvement)

The increase in test R^2 suggests better generalization, but the drop in training R^2 indicates potential underfitting due to the limited generality of the keywords.

Feature Insights

Key predictors included:

- Budget: The strongest driver of box office revenue.
- Genre: "Action," "Adventure," and "Fantasy" were associated with higher revenue, while "Drama" and "Documentary" had less impact.
- Keywords: Terms like "universe" and "strength" showed positive contributions, particularly for blockbuster genres.

Seasonality and Outliers

Release timing also played a significant role:

- Holiday Season (Nov–Dec): Highest average revenue, driven by increased audience engagement and marketing.
- Summer Blockbusters (Jun–Aug): Consistently strong performers due to traditional blockbuster timing.
- Other Seasons: Films released outside these periods saw lower average revenues.

Outliers like Avatar and Titanic heavily influenced the dataset. While the keyword-augmented model improved predictions for these films, unique, high-grossing movies remained difficult to model.

Limitations of Keywords

Many high-revenue keywords were tied to specific outlier films, such as Titanic or Transformers, limiting their broader applicability. This underscores the need for a more comprehensive and flexible approach to incorporating keywords into predictive models.

Conclusion

To answer the question of whether genre-specific keywords enhance the predictive accuracy of a box office success model, we reached ambiguous conclusions about the effects of keywords. Comparing a baseline model —built on features such as genre, studio, budget, and season of release— and a keyword-enhanced model, we found a lower evaluation metric for the baseline test data (R^2= 0.586) compared to the keyword-enhanced test data (R^2 = 0.634). The keyword-enhanced model achieved a higher R-Squared than the baseline model, suggesting that the keyword model explains a greater amount of variance in box office performance on unseen data. Additionally, a lower test RMSE in the keyword-enhanced model (RMSE= 108.08 million) indicated a more accurate average prediction of revenue in comparison to the baseline model (RMSE= 97.60 million). These positive changes in R-Squared and RMSE in the unseen data of the keyword-enhanced model suggest that incorporating specific keywords adds predictive value beyond features like genre, studio, budget, and release date. However, using only a few keywords in our model, due to many film-specific rather than genre-specific keywords in our data, limited our ability to draw conclusive results about the impact of keywords on box office success.

Assessing the role of specific keywords in predicting box office success is a challenging task due to inherent biases. For instance, while prequels and sequels can be significant indicators of success, their profitability is often rooted in the franchise's prior success and the audience's familiarity, rather than the keywords themselves. On the other hand, standalone films with distinctive keywords add another layer of difficulty. For example, take the iconic film *Titanic*. Does the keyword “Titanic” actually predict box office success, or is it the movie’s success that gives the keyword its predictive power? This raises the common issue of causality, as we need to

make sure that the association flows from keyword to box-office success, and not the other way around, in order to come to meaningful conclusions.

Apart from the limited keyword usage in our model and the challenge of determining a clear association between keywords and revenue, unaccounted interactions between keywords and other features in our model could explain our results. These interactions are hard to observe because the keyword variable is multi-faceted and high dimensional. In our analysis, we found that words with high frequencies were often film specific,while lower frequency words were more generalizable. This led to the decision of using lower frequency but more generalizable words to improve the model's application across films. However, the trade-off between generalizability and frequency is likely the impeded ability to capture all the nuances present in our data. Without the impact of the frequent yet content specific keywords and information in our data we might not have been able to observe the full extent keywords have on predictive accuracy. However, while the predictive accuracy of keywords in our model remains ambiguous, clearer results could be achieved with a dataset where the keywords are more general and consistently applicable across films.

In future research, standalone movies and series (those with prequels or sequels) could be analyzed separately to address the challenges of bias and causality in our keyword analysis. For standalone movies, removing the movie's name from the keyword list would eliminate issues where the name itself influences the predictive power, ensuring a clearer focus on other meaningful factors. For series, we could evaluate whether the original movie or the average performance of its sequels better predicts box office success. If combining standalone and series

movies in the same analysis, only the original movie (e.g., Shrek, Harry Potter and the Philosopher's Stone, The Lion King) should be included to avoid overrepresentation and maintain consistency. These adjustments would provide a more balanced and accurate analysis of keyword impact.

References

"Coming to a Screen Near You: How Data Science Is Revolutionizing the Film Industry." DataColumn, North Carolina State University, 30 Jan. 2023, https://datacolumn.iaa.ncsu.edu/blog/2023/01/30/coming-to-a-screen-near-you-how-data-science-is-revolutionizing-the-film-industry/.

Utkarshx27. *Movies Dataset*. Kaggle, 01 October 2024, https://www.kaggle.com/datasets/utkarshx27/movies-dataset

Yoo, Youngjae, Justin Kanter, and Ian Cummings. "Predicting Movie Revenues Using IMDb Data." *CS229 Machine Learning Projects*, Stanford University, 2011, https://cs229.stanford.edu/proj2011/YooKanterCummings-PredictingMovieRevenuesUsingImdbData.pdf.