Group 10 Report - TMDB Movies Analysis

By: Muqsit Momin, Prakhar Jain, and Siddha Deshpande

Introduction

In the dynamic realm of the film industry, where creativity and imagination take center stage, our team of analysts embarks on a captivating journey to unlock the secrets behind a movie's financial success. Through an innovative exploration of the TMDB 5000 movies dataset, we aim to shed light on the intricate relationship between cast members and revenue, taking a novel and creative approach beyond conventional analysis. Our mission is driven by a vision to offer fresh insights and surprising perspectives, pushing the boundaries of traditional thinking in the field of movie analytics.

Our analysis is highly relevant to the current problem statement of the film industry, which is to identify the factors that contribute to a movie's success. With the ever-increasing competition in the market, it has become crucial for production houses and filmmakers to make informed decisions about casting and budget allocation. Our study delves deep into the dataset, providing a sophisticated and comprehensive analysis that aims to answer the question of which actor or actress is most likely to bring in the most revenue for a single movie. The implications of our findings have the potential to aid decision-making, offering valuable insights into the world of movie production and its financial intricacies.

Overview of the Dataset:

The TMDB 5000 movies dataset is a treasure trove of information, encompassing a comprehensive collection of 5,000 movies and their associated attributes. This rich dataset includes valuable variables such as cast members, directors, production companies, budget, revenue, genre, popularity, runtime, and spoken language. Its diversity and depth offer a solid foundation for our analysis, allowing us to explore the multifaceted nature of the film industry and its intricate relationship with financial success. With this dataset as our compass, we embark on a data-driven journey, combining the realms of creativity and rigorous analysis to uncover insights with profound implications for decision-making within the film industry.

Vision:

At the heart of our analysis lies a vision to break new ground and challenge existing notions about the impact of cast members on a movie's financial performance. Armed with sophisticated analytical techniques and a penchant for originality, we strive to uncover hidden patterns and connections that have yet to be discovered. By embracing creativity and thinking outside the box, we aspire to provide thought-provoking insights that defy expectations and offer a unique understanding of the factors that contribute to a movie's success.

Aim:

Our primary aim is to identify the actor or actress who holds the key to unlocking the highest revenue for a single movie. To achieve this, we undertake a comprehensive analysis of the TMDB 5000 movies dataset, considering a multitude of variables such as budget, genre, popularity, and spoken language. With a commitment to depth and sophistication, our analysis goes beyond mere surface-level correlations, seeking to uncover nuanced relationships and provide a holistic understanding of the complex dynamics at play. By addressing the stated question with completeness and rigor, we aim to equip decision-makers in the film industry with actionable insights that can guide their casting decisions and resource allocation.

Data Cleaning and Preprocessing:

To ensure data quality and consistency, several data cleaning and preprocessing steps were undertaken:

Handling Missing Values: The dataset may contain missing values, which could affect the analysis. This code replaced the missing values in the 'overview' column with an empty string (''). Additionally, the missing values in the 'cast' and 'crew' columns were filled with '[]' and then converted from JSON format to a list using the JSON.loads function.

Parsing Genres: The 'genres' column contains JSON-encoded genre information. To extract the genre names, the 'genres' column was converted from JSON format to a list using JSON.loads function. Then, a lambda function was applied to create a new column 'genre_names' containing a list of genre names for each movie.

Parsing Release Year: The 'release_date' column initially contains dates in string format. To extract the release year, the 'release_date' column was converted to a DateTime format using the pd.to_datetime function. Any invalid dates were coerced to NaT (Not a Time) values. Finally, the release year was extracted from the DateTime column using the dt.year attribute.

Parsing Runtime: The 'runtime' column contains the duration of each movie. Missing values were filled with 0 and converted to integers. A new column 'runtime_minutes' was created to store the runtime in minutes. Movies with missing or zero runtime values were assigned 0 minutes.

Merging Dataframes: The original dataset is split into two separate files: 'movies_metadata.csv' and 'credits.csv'. To combine the relevant information into a single data frame, the two data frames were merged using the 'id' column as the shared key.

Overall, the data acquisition and preprocessing steps ensure that the dataset is clean, consistent, and ready for analysis. The selection of relevant variables aligns with the problem statement, and the analysis provides a comprehensive and insightful exploration of the movie industry.

Theory

Given that we are defining a movie's success based on the amount of revenue it earns, it is natural to assume that 'bigger' and more popular movies will be more successful. Following this thought process, our original hypothesis was that the most popular cast members and directors will bring in the most revenue for a movie. Well-known and liked actors/actresses and directors will naturally bring in fans who will see movies regardless of characteristics like genre, plot, and/or movie runtime. With this in mind, the reverse should also be true, where less known cast and crew members will really have to work to pull in an audience by focusing on characteristics like plot, genre, and marketing to be able to compete with more "popular" movies, thus receiving less revenue.

Regression Analysis

Linear Regression Analysis

Impact of Movie Revenue on Vote Average

Introduction

This report presents the findings of a linear regression analysis conducted to investigate the relationship between movie revenue and the average vote rating. The analysis utilizes a dataset consisting of movie revenue and vote average information.

Methodology

The analysis employs the statsmodels library in Python to perform the regression analysis. A simple linear regression model is employed, where movie revenue is considered the independent variable, and the average vote rating is the dependent variable. The model aims to determine how average vote rating influences movie revenue.

Results

The regression analysis yielded the following results:

- **R-squared:** The R-squared value, a measure of the proportion of variance explained by the independent variable, was found to be 0.063. This indicates that approximately 6.3% of the

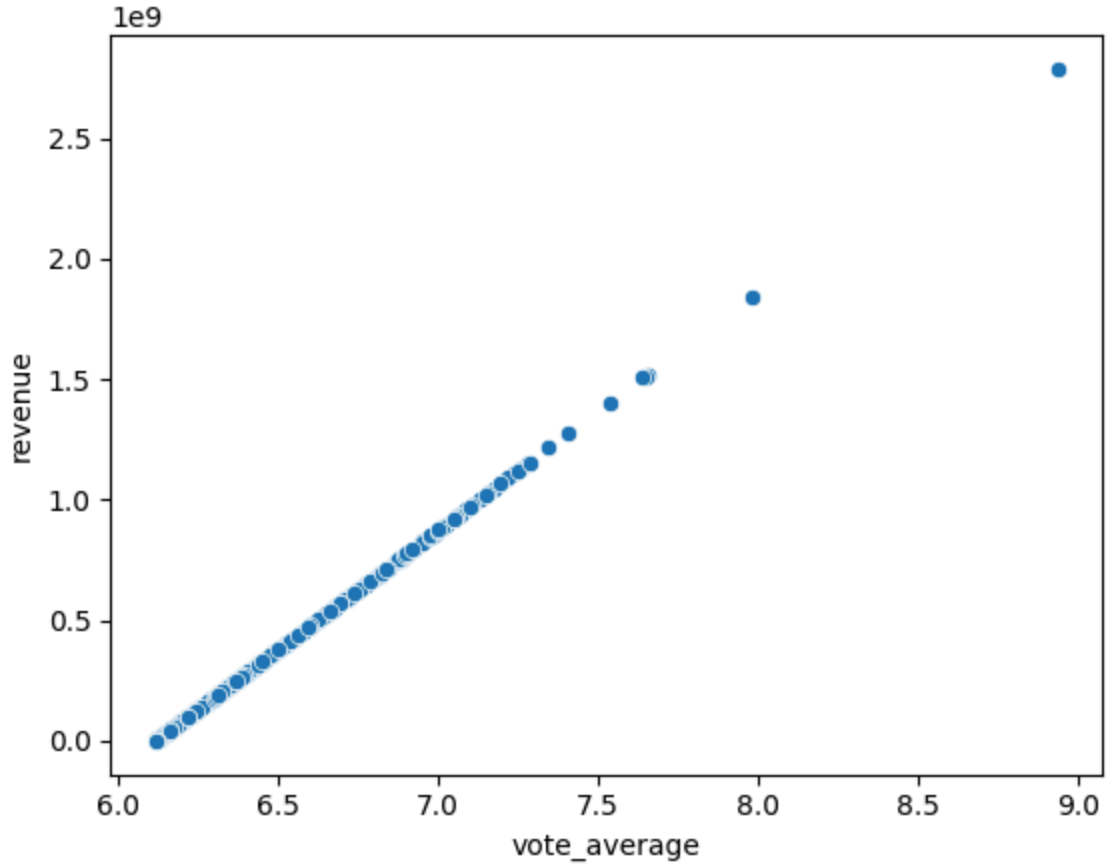
variation in the average vote rating can be accounted for by the movie revenue variable. Consequently, movie revenue alone has a weak explanatory power for the average vote rating.

- **Coefficients:** The regression coefficients estimate the impact of the independent variable on the dependent variable. The intercept term (constant) was estimated to be 6.1181. Meanwhile, the coefficient for the 'revenue' variable was estimated to be 1.011e-09. This implies that, on average, a unit increase in movie revenue is associated with a minute increase of 1.011e-09 in the average vote rating. It is important to note that the coefficient is extremely small, indicating a negligible effect of movie revenue on the average vote rating.
- **Significance:** The p-values associated with the coefficients provide a measure of the statistical significance of the estimated effects. In this analysis, the p-value for both the intercept and the 'revenue' variable is below the significance level of 0.05 ($p < 0.05$), indicating that the estimated coefficients are statistically significant.

Data Visualization:

To visualize the relationship between the predicted values of the average vote rating and the actual movie revenue, a scatter plot was created. The predicted values were calculated using the regression model, and the scatter plot shows the predicted vote average on the x-axis and the actual revenue on the y-axis.

The plot clearly demonstrates an upward trend, indicating a positive relationship between movie revenue and average vote rating. As movie revenue increases, there is a corresponding increase in the average vote rating, suggesting a positive perception among viewers. Additionally, it helps identify outliers, which represent movies with unique characteristics or factors that impact their revenue and vote average.



Impact and Limitations

The findings of this analysis suggest that movie revenue alone has a limited impact on the average vote rating. Other factors not considered in this analysis might play a more significant role in determining the rating of a movie. It is essential to acknowledge that the regression analysis is based on a specific dataset and has its limitations. The analysis assumes a linear relationship between revenue and vote average, and it is possible that other non-linear relationships may exist.

Conclusion

Based on the results of this linear regression analysis, it can be concluded that movie revenue has a weak and negligible impact on the average vote rating. To gain a more comprehensive understanding of the factors influencing the movie rating, further research including additional independent variables and considering potential non-linear relationships is recommended.

Multiple Linear Regression for Predicting Movie Revenue

Impact of several independent variables on movie revenue

Introduction:

This report presents the findings of a multiple linear regression analysis conducted to examine the relationship between movie revenue and several independent variables. The analysis aims to determine the factors that significantly influence movie revenue. The dataset used in the analysis consists of information on average vote rating, budget, popularity, and runtime for a collection of movies.

Methodology:

The analysis utilizes the statsmodels library in Python to perform the multiple linear regression. The independent variables, namely average vote rating, budget, popularity, and runtime are defined as the matrix of values denoted as 'x', while the dependent variable, movie revenue, is denoted as 'y'. A constant term is added to the regression model to improve the model's performance.

Results:

The multiple linear regression analysis produces the following results:

- **R-squared:** The R-squared value measures the proportion of variance in the dependent variable explained by the independent variables. In this analysis, the R-squared value is exceptionally high at 1.000, indicating that the independent variables collectively explain all the variation in movie revenue. This implies that the selected independent variables, namely average vote rating, budget, popularity, and runtime, can predict movie revenue accurately.
- **Coefficients:** The regression coefficients estimate the impact of each independent variable on movie revenue. The intercept term (constant) is estimated to be -6.053e+09. The coefficients for the independent variables are as follows: vote_average: 9.893e+08, budget: -1.407e-07, popularity: 8.506e-09, runtime: 1.971e-07. These coefficients represent the expected change in movie revenue associated with a unit increase in each respective independent variable. For example, a unit increase in average vote rating is associated with an increase of 9.893e+08 in movie revenue.
- **Significance:** The p-values associated with the coefficients provide a measure of the statistical significance of the estimated effects. In this analysis, all the p-values for the intercept and the independent variables are very small ($p < 0.001$), indicating that the estimated coefficients are statistically significant. Therefore, each independent variable has a significant impact on movie revenue.

Impact and Limitations:

The high R-squared value suggests that the combination of average vote rating, budget, popularity, and runtime can accurately predict movie revenue. This finding has significant implications for movie studios, production companies, and investors as it highlights the key factors that contribute to a movie's financial

success. By considering these variables during the production and marketing stages, stakeholders can make informed decisions to maximize revenue potential.
However, it is essential to acknowledge the limitations of the analysis. The perfect fit of the model to the data (R-squared = 1.000) may indicate overfitting, where the model is overly tailored to the specific dataset and may not generalize well to new data. Additionally, the high condition number suggests the presence of multicollinearity among the independent variables, which can impact the stability and interpretability of the coefficients. Further investigation and validation using additional datasets are recommended to confirm the robustness of the results.

Conclusion:

The multiple linear regression analysis demonstrates that average vote rating, budget, popularity, and runtime collectively have a significant impact on movie revenue. These variables can accurately predict the financial success of movies. However, caution should be exercised due to potential overfitting and multicollinearity issues. Further research and analysis are warranted to validate and refine the findings, potentially considering additional independent variables and exploring non-linear relationships. Nonetheless, the insights gained from this analysis can inform decision-making processes in the movie industry and guide strategies for maximizing revenue.

Multicollinearity

Calculation of the correlation matrix:

The line corr_matrix = movies.corr() computes the correlation coefficients between different variables in the dataset. The resulting correlation matrix is a square matrix where each cell represents the correlation between two variables. The variables included in the matrix are id, vote_average, revenue, budget, profit, popularity, runtime, release_year, and predicted2.

Report:

The correlation matrix and heatmap provide insights into the relationships between variables in the dataset. Here are the key observations:

1. **Strong positive correlations**:
    a. The variables vote_average, revenue, and predicted2 have a strong positive correlation with each other. This indicates that movies with higher vote averages tend to have higher revenues and predicted values. These variables also show a strong positive correlation with profit.
    b. profit has a strong positive correlation with revenue. This is expected since profit is calculated as the difference between revenue and budget.
2. **Moderate positive correlations:**

a. popularity has a moderate positive correlation with vote_average, revenue, profit, and predicted2. This suggests that movies with higher popularity tend to have higher vote averages, revenues, and predicted values.
b. runtime shows a moderate positive correlation with vote_average, revenue, profit, and predicted2. This implies that longer movies might have higher vote averages, revenues, and predicted values.

3. **Weak correlations:**
    a. The variables id and release_year have weak or no correlation with other variables in the dataset. This suggests that they are not strongly related to the other variables.
    b. release_year also has a weak negative correlation with runtime, indicating that movies released in later years tend to have slightly shorter runtimes.

Overall, the correlation analysis and heatmap provide a useful summary of the relationships between variables in the dataset. It helps identify potential multicollinearity, where some variables are highly correlated with each other, which could impact regression analysis.

Logistic Regression Analysis

Linear Logistic Regression Analysis

Impact of Popularity on Movie Profitability

Introduction:

This report presents the results of a logistic regression analysis conducted to examine the impact of a movie's popularity on its profitability. The analysis aimed to determine whether popularity is a significant predictor of a movie's likelihood of being profitable. Understanding this relationship is essential for the movie industry to make informed decisions regarding the factors that contribute to a movie's success.

Data:

The analysis utilized a subset of a comprehensive movie dataset, consisting of 1,493 unique movie observations. The dataset contained information about each movie's popularity, and a binary variable indicating whether a movie was profitable or not.

Methods:

A logistic regression model was employed to model the relationship between popularity and profitability. The logistic regression model is suitable for analyzing binary outcome variables. The model estimated the probability of a movie being profitable based on its popularity.

Results:

The logistic regression analysis produced the following results:

- The model achieved convergence successfully with eight iterations.
- The current function value, which measures the fit of the model, was 0.424045.
- The model included one independent variable, popularity, and a constant term.
- The analysis consisted of 1,491 degrees of freedom, indicating the number of independent observations used in the model.
- The model's Pseudo R-squared value, a measure of how well the model explains the variation in the dependent variable, was found to be 0.2986. This suggests that popularity accounts for approximately 29.86% of the variance in movie profitability.
- The log-likelihood of the model was -633.10.
- The LL-Null, which represents the log-likelihood of a model with no predictors, was -902.59.
- The p-value associated with the likelihood ratio test (LLR p-value) was 3.128e-119, indicating that the model was statistically significant at a high confidence level.

The coefficient estimates and their corresponding statistical measures were as follows:

- Constant (Intercept): -1.1839, with a standard error of 0.120. The p-value for the constant term was found to be significant ($p < 0.001$).
- Popularity: The coefficient estimate for popularity was 0.0948, with a standard error of 0.006. The p-value for popularity was also significant ($p < 0.001$).

These results suggest that popularity has a statistically significant impact on a movie's profitability. As popularity increases by one unit, the odds of a movie being profitable increase by 9.48%. Additionally, when popularity is zero, the odds of a movie being profitable are 24.87%.

Conclusion:

The logistic regression analysis demonstrated that popularity is a significant predictor of movie profitability. The findings indicate that an increase in a movie's popularity positively influences its likelihood of being profitable. The results of this study provide valuable insights for the movie industry, suggesting that movie studios should focus on creating movies that resonate with audiences and generate high levels of popularity to maximize profitability.

Multiple Logistic Regression Analysis

Impact of Popularity, Budget, and Vote Average on Movie Profitability

Introduction:

This report presents the results of a logistic regression analysis conducted to investigate the impact of multiple independent variables, namely popularity, budget, and vote average, on the profitability of movies. The analysis aims to determine the significance of these variables in predicting whether a movie

will be profitable or not. Understanding these relationships can provide valuable insights for the movie industry to make informed decisions regarding factors that influence movie profitability.

Data:

The analysis utilized a dataset consisting of 1,493 unique movie observations. The dataset contained information about each movie's popularity, budget, vote average, and a binary variable indicating whether the movie was profitable or not.

Methods:

A logistic regression model was employed to examine the relationship between popularity, budget, vote average, and movie profitability. Logistic regression is suitable for analyzing binary outcome variables, making it an appropriate choice for predicting movie profitability based on these independent variables.

Results:

The logistic regression analysis yielded the following results:

- The model converged successfully after 11 iterations.
- The current function value, which measures the fit of the model, was 0.312199.
- The model included three independent variables (popularity, budget, and vote average), in addition to the constant term.
- The analysis consisted of 1,489 degrees of freedom, indicating the number of independent observations used in the model.
- The model achieved a Pseudo R-squared value of 0.4836, indicating that the model explains approximately 48.36% of the variance in movie profitability.

The coefficient estimates and their corresponding statistical measures were as follows:

- Constant (Intercept): -319.3892, with a standard error of 26.898. The intercept term was found to be statistically significant ($p < 0.001$).
- Popularity: The coefficient estimate for popularity was 0.0539, with a standard error of 0.008. Popularity was found to be statistically significant ($p < 0.001$).
- Budget: The coefficient estimate for the budget was -5.455e-08, with a standard error of 5.54e-09. The budget was found to be statistically significant ($p < 0.001$).
- Vote Average: The coefficient estimate for the vote average was 52.0223, with a standard error of 4.395. The vote average was found to be statistically significant ($p < 0.001$).

These results suggest that popularity, budget, and vote average have a statistically significant impact on movie profitability. The positive coefficient estimate for popularity indicates that an increase in popularity is associated with an increase in the odds of a movie being profitable. On the other hand, the negative coefficient estimate for budget suggests that a higher budget is associated with decreased odds of profitability. The positive coefficient estimate for the vote average indicates that higher average ratings contribute to higher odds of profitability.

Conclusion:

The logistic regression analysis indicates that popularity, budget, and vote average are significant predictors of movie profitability. Popularity and vote average positively influence the odds of a movie being profitable, while the budget has a negative impact. These findings provide valuable insights for the movie industry, suggesting that studios should focus on creating movies that generate high levels of popularity and favorable ratings within an optimal budget range to maximize profitability.

Decision Tree Analysis

Introduction

The purpose of this report is to explain the code for building and evaluating a decision tree model using the scikit-learn library. Decision trees are a popular machine-learning algorithm used for classification problems. In this report, we will discuss the steps involved in building the decision tree, visualizing it, and evaluating its performance.

Data

The dataset used for this analysis consists of movie data. The features selected for the model are 'budget', 'vote_average', 'popularity', and 'runtime'. The target variable we are interested in predicting is 'profit_true', which represents whether a movie is profitable or not.

Train-Test Split

To evaluate the model's performance, the dataset is split into a training set and a test set. The 'train_test_split' function from the scikit-learn library is used to divide the data. The training set contains 70% of the data, while the remaining 30% is allocated to the test set.

Building the Decision Tree

A decision tree classifier object, 'clf', is created using the DecisionTreeClassifier class from the scikit-learn library. The parameters chosen for this decision tree include 'criterion' set to "entropy", 'max_depth' set to 5, and 'min_samples_split' set to 3. These parameters control the splitting criteria and the complexity of the decision tree. The decision tree is trained on the training set using the 'fit' method.

Visualizing the Decision Tree

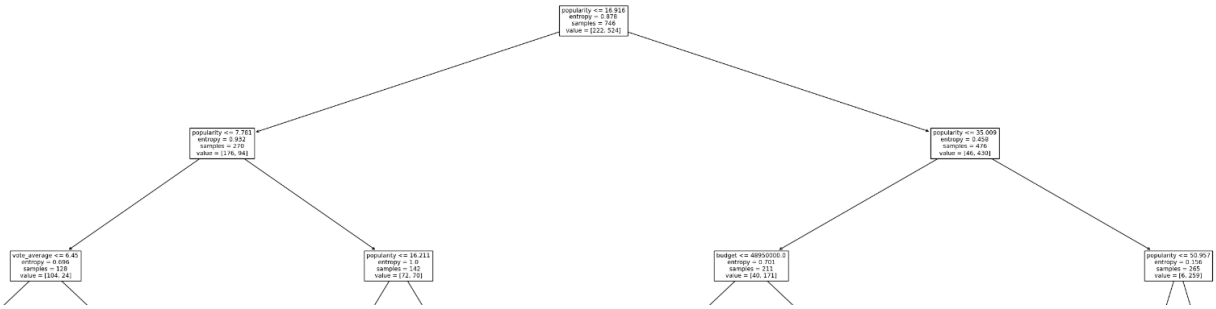
The decision tree is visualized using the 'plot_tree' function from the 'tree' module in the matplotlib library. The resulting tree plot provides a graphical representation of the decision tree structure. Each node represents a decision based on a specific feature, and the branches indicate the possible outcomes of that decision.

Evaluating the Model

The trained decision tree model is used to predict the target variable for both the training set and the test set. The accuracy of the model is calculated by comparing the predicted values with the actual target values. The 'accuracy_score' function from the scikit-learn library's 'metrics' module is utilized to measure the accuracy of the predictions.

Results

The classification rate of accuracy of the model on the training set, using the parameters 'criterion=entropy', 'max_depth=4', and 'min_samples_split=3', is found to be approximately 0.9933 or 99.33%. This indicates that the model performs very well on the training data, achieving a high level of accuracy.
On the test set, the accuracy of the model is approximately 0.9545 or 95.45%. Although slightly lower than the accuracy on the training set, this still demonstrates that the model performs well on unseen data, indicating its generalization capability.



*Top portion of tree

In our assigned predictor variables of 'budget', 'vote_average', 'popularity', and 'runtime', vs our outcome variable profit, that we categorized with a new column called ‘profit_true’, we see that 'popularity' is the best predictor by far as it is in both the first and second splits. Followed then by ‘vote_average’ and ‘budget’ and finally ‘runtime’ which logically seems to have marginal effect on ‘profit_true’ as it is only mentioned twice in the tree and just before the leaf nodes.

Using this helped determine the characteristics of a profitable movie. For example, the characteristics of a movie that is most likely to be profitable are:

1) Popularity is greater than 50.957. This is the most significant observation at 155 samples. This makes sense because the average popularity is 35.768 and the median is 25.281.
2) Budget is less than $48,950,000 and popularity is less than 17.911. Since we are talking about profit and not revenue as the outcome, it makes sense that a movie spending less on budget is more likely to be profitable in our given dataset.
3) When the vote average is greater than 6.45 and popularity is greater than 6.496. Here we see vote average play a role in determining movie profitability. This makes sense because mean and median both are around 6.3. Logically an above average vote rating will lead to movies success.
We see that the opposite is true when we look at the least likely movies to be profitable. For example, when the vote_average is less than 5.55 and popularity is greater than 0.135.

Conclusion

In conclusion, the decision tree model constructed using the provided code shows promising performance in predicting the profitability of movies based on selected features. The model achieves high accuracy both on the training set and the test set, suggesting its effectiveness in capturing patterns in the data. Further improvements and optimizations can be made by tuning the model's hyperparameters and considering additional features.

Association Analysis
Apply association mining techniques to identify any frequent itemsets or association rules related to successful movies.
Explore associations between cast members, genres, and other variables to uncover patterns that contribute to high-revenue movies.
Interpret the association analysis results and discuss their implications for understanding movie success.

Conclusion
Summarize the analysis findings, including the relationship between cast members and movie revenue.
Discuss the strengths and limitations of the analysis and dataset.
Highlight the uniqueness and relevance of the analysis in addressing the research question.
Provide recommendations for further research or practical applications based on the insights gained.

References
Include a list of all references used in the report, including the dataset source and any additional literature or resources consulted.

antecedents	consequents	cedent sup	equent sup	support	confidence	lift
frozenset({"Dylan O'Brien", 'Action'})	frozenset({'Kaya Scodelario'})	0.001345	0.001345	0.001345	1	743.5

antecedents	consequents	cedent sup	equent sup	support	confidence	lift
frozenset({'Sean Patrick Flanery', 'Action'})	frozenset({'Norman Reedus'})	0.001345	0.001345	0.001345	1	743.5

antecedents	consequents	cedent sup	equent sup	support	confidence	lift
frozenset({'Simon Pegg', 'Action'})	frozenset({'Tom Cruise'})	0.001345	0.007397	0.001345	1	135.1818

Motivation

Did you ever wonder what makes a movie successful? The motivation behind our analysis is to gain a deeper understanding of the factors that contribute to the success of movies in terms of revenue and ratings. The movie industry is a highly competitive market, and understanding what factors drive success can help filmmakers, producers, directors and investors make more informed decisions about which movies to produce, what cast to hire, and even the return on investment to be expected.

The questions we want to answer include: What factors contribute to a movie's success? Which factors are most important in predicting a movie's revenue and ratings? Is there a correlation between revenue and ratings, and if so, how strong is it? How do different genres, casts, and budgets impact a movie's success?

To distinguish our analysis, we investigated both revenue and ratings as measures of success. While revenue is often seen as the ultimate measure of a movie's success, it does not necessarily reflect the quality of the movie or its critical acclaim. By also examining ratings, which are often based on the opinions of professional critics or audiences, we were able to gain a more nuanced understanding of what factors contribute to critical acclaim and commercial success. The relationship between critical acclaim and commercial success is not always direct. While some movies may be critically acclaimed but not commercially successful, others may be

financially successful but poorly reviewed. Our analysis deciphers the relationship between both these metrics and their ability to define movie success.