The Crystal Ball of Cinema: Using Data to Predict Movie Success

Albert Sanz
November 2023

Outline

- Executive Summary
- Introduction
- Data Collection and Preprocessing
- Exploratory Data Analysis (EDA)
- Predictive Modelling
- Results and Findings
- Limitations and Conclusions
- References

Executive Summary

- **The project's objective is to support early-stage movie investment decisions by employing machine learning models to forecast a movie's success. This is defined as a Return On Investment (ROI) equal or greater than 2, equivalent to grossing 3 times the production budget at the global box office.**
- Data was extracted from The Movie Database (TMDB) API and using web scraping techniques. Exceptionally high ROI values (exceeding 40) were excluded from the dataset as they represent a unique subset of highly successful, low-budget movies.
- Initially comprising 3,187 movies, our dataset was refined to **1,490 movies** after handling missing values. Feature engineering and data transformation techniques were applied to create meaningful variables and prepare the data for modelling. Exploratory Data Analysis provided insights into features and relationships and dependencies among them.
- Some of the most important binary classification algorithms and ensemble methods were implemented, including Logistic Regression, K Nearest Neighbours (KNN), Support Vector Machines (SVM), Decision Tree, Random Forest, Extreme Gradient Boosting (XGBoost) and Voting Classifiers (Soft and Hard).
- Among these models, **XGBoost** displayed superior predictive performance metrics compared to standalone models, offering slightly lower overall performance than ensemble-based voting classifiers. However, its interpretability, scalability, explicability, and ease of maintenance make it **our preferred choice for predicting movie success**.
- Cross-validation and robust evaluation techniques were implemented to ensure the model's generalizability and reliability on unseen data. The model was also adjusted to mitigate class imbalance and give more weight to successful movies, improving minority class prediction.
- Our model, the Movie Success Predictor (MSP), enables **dynamic adjustment of decision thresholds to accommodate diverse user requirements**, facilitating tailored decision-making.
- While MSP showed promising results, it exhibited a **heavy reliance on the "Sequel" feature**. Future iterations will focus on Feature Engineering, Selection, and Dimensionality Reduction techniques to diversify feature importance and reduce dependency on this singular attribute.
- The model demonstrated reasonable performance metrics: Accuracy of 70%, Precision of 61%, Recall of 48%, ROC AUC of 65%, and PR AUC of 48%. Notably, the **model's precision rate of 61% marks a significant improvement compared to** the industry's success rates, which stand at **29% across all studios and 41% among the top 50 studios** producing a higher volume of movies.
- MSP recommended investment in 127 out of 447 movies in the testing set, projecting $58.2B worldwide box office against a $13.2B budget with a **Total ROI of 3.4**.
- The **Average ROI, at 4.1 and within the 89th percentile**, showcases MSP's capability among studios producing more than five movies over the last two decades. Only around 10 out of the 110 studios we investigated delivered a better Average ROI than our model.
- Given the movie industry's dynamic nature, continuous refinement of predictive models is essential. Collaboration with industry professionals for feedback and model adaptation remains vital for sustaining accuracy.

Introduction: Project Overview

The Significance of the Film Industry : A Box Office Perspective

The film industry stands as a global centre of entertainment, captivating audiences worldwide while contributing significantly to the economy. One of the most compelling indicators of its influence lies in the immense box office revenues generated annually. Research firm Gower Street Analytics calculated that global box office reached an **estimated $26 billion total** in 2022, highlighting cinema's universal appeal and cultural impact.

Navigating Financial Realities in Filmmaking

Beyond the glamour of the silver screen lies an intricate financial landscape, with movie **production budgets often ranging between $100 and $200 million**. These costs encompass a multitude of factors, including casting, script development, filming, post-production, marketing, and distribution. Collectively, these expenses reflect the substantial investment needed to bring a movie to fruition.

Challenges of Financial Profitability

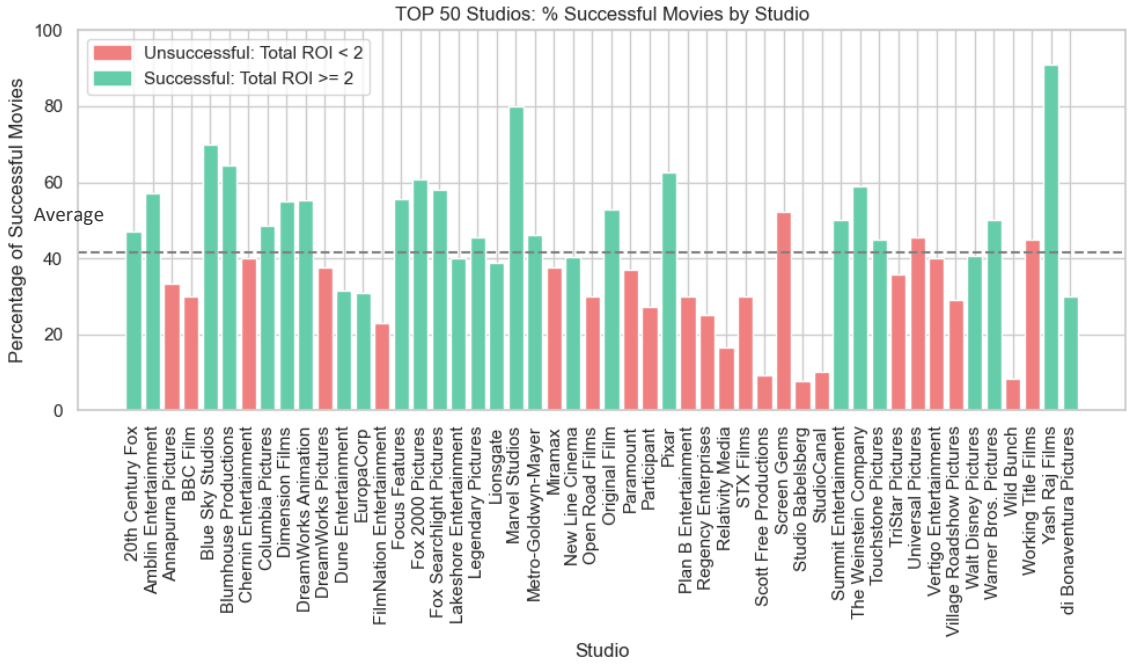
Achieving financial profitability in the movie industry remains a persistent challenge. Factors like audience preferences, market competitiveness, evolving consumption trends, and the inherent risk in creative endeavours contribute to this complex scenario.

Unexpected Setbacks and Industry Uncertainty

Many movies, once anticipated as box office triumphs, ended up being a disappointment. The 2019 film "Cats," an adaptation of Andrew Lloyd Webber's renowned musical that featured a star-studded cast, faced a commercial disaster, resulting in financial losses estimated at $100 million. This instance underlines the **industry's uncertainty**, where seemingly promising ventures encounter **substantial setbacks**. However, the industry has also witnessed **unexpected triumphs** and innovative strategies that have defied conventional norms, leading to remarkable financial returns.

Lights, Camera, Prediction: Harnessing Data for Movie Success

The chart provided illustrates that **among the top 50 studios** producing a higher volume of movies, the **average proportion of financially successful films stands at approximately 40%**. However, **when considering all movie producers, this figure drops to 29%**. Furthermore, only 27 out of these 50 studios exhibit a Total Return On Investment (Total ROI) exceeding 2. This data strongly indicates that implementing predictive modelling techniques could notably enhance studios' abilities to predict and potentially increase the probability of generating financially successful movies prior to their release.



Introduction: Problem Statement

Defining movie success

Movie success encompasses various metrics, including popularity, critical acclaim, and box office revenue. However, for the purpose of this project, our primary focus revolves around the metric of return on investment (ROI), representing the profitability per dollar invested. The key reason for **using ROI as a target metric** is that it explains the financial success of a movie, taking into account both costs and revenue. This metric also reflects how efficiently a studio leverages its assets to generate profits, emphasizing the need for optimal investment utilisation.

Hollywood's financial complexity and Revenue-Cost Ambiguity

The film industry operates within a labyrinth of financial intricacies. **Enormous marketing and distribution expenditures** often rival the production budget, complicating the assessment of a movie's true costs. Determining the authentic production budget is equally challenging due to **creative accounting practices** aimed at manipulating figures, often inflating budgets to avoid profit participation payouts to key creatives.

Additionally, the calculation of total revenues encounters ambiguity as movies earn income from diverse avenues beyond theatrical releases. While theatrical releases significantly contribute to revenue, studios typically receive a fraction of the box office gross (about 50% domestically and 40% internationally). The complexities intensify as movies generate revenue from various platforms such as airlines, hospitality, home entertainment, Video on Demand (VOD), television, and merchandising, making it **challenging to ascertain the complete revenue picture**.

Establishing a Benchmark for Success

A rule of thumb often used in the film industry is known as the "2.5x Rule." According to this guideline, a movie needs to gross around 2.5 times its production budget at the global box office to be considered minimally profitable. We raised the bar of profitability for a movie to avoid uninteresting returns and **defined success as a ROI equal or greater than 2**, which is equivalent to grossing 3 times the production budget at the global box office. A ROI of 2 would be around the 65th percentile in our dataset, which entails a movie be considered successful if its ROI is within top 35% of all movies. Setting a higher ROI can be justified through the following arguments:

1) **Risk mitigation**: Ensures a buffer against unexpected costs or underperformance. It also safeguards against market volatility and uncertainty.

2) **Long-term viability**: A higher ROI provides better financial stability and supports future projects. It allows for reinvestment into new ventures and overall growth for the production company.

3) **Brand building and investor confidence**: Consistently achieving a higher ROI can enhance the reputation of the production company and individuals involved, attracting more opportunities and partnerships in the future. It may also attract more investors and increase their confidence in the project's potential success

Objective

This project aims to aid early-stage movie investment decisions by utilizing machine learning models to predict movie's profitability, assisting studios in making informed choices

Data Collection and Preprocessing

Data Collection: Sources

The Movie Database (TMDB) API

The analysis relies on data extracted from **The Movie Database (TMDB) API** as of October 2023. It's important to note that while the analysis utilizes TMDB API data, it is not endorsed or certified by TMDB. TMDB serves as a widely used community-built movie and TV database, accumulating data contributed by its community since 2008.

The project focuses on **movies released between 2003 and 2022**, encompassing 20 years to reflect the industry's current landscape. To ensure revenue accuracy, 2023 was excluded from the data. The dataset specifically includes **movies released in Great Britain**, excluding those with adult content. These criteria narrowed the dataset to 42,796 movies, averaging approximately 2,100 movies per year.

The decision to focus on movies released in Great Britain was strategic, aiming to **balance dataset size reduction while ensuring global impact**. The Great Britain film market acts as a crucial global hub, often hosting movies with widespread international reach. By focusing on movies released in this market, the analysis encompasses films that possess qualities transcending national boundaries and resonate with diverse global audiences. This approach ensures a curated dataset that captures movies with the potential for international success, facilitating a more focused analysis while retaining the capacity to study factors influencing financial success in the broader context of the global film industry.

From this initial dataset, a subset of c.10,000 movies was selected based on popularity, sorted in descending order. Popularity is determined by TDMB based on various attributes such as daily views and votes, user favourites, and watchlist additions. This method aimed to **prioritise movies likely to contain essential information** for the project's analysis as TMDB members are more inclined to contribute information for popular and well-known movies. Notably, an examination of the data revealed that 84% of the top 1,000 most popular movies provided all required data (e.g., Revenue or Profit), while only 3% of movies ranked between 9,000 and 10,000 contained such essential information.

Web scraping

The project also utilises **web scraping techniques** to gather two additional features: the average worldwide annual box office revenue and Academy Awards for best actors. The box office revenue data was sourced from Box Office Mojo, renowned for its box office statistics, while information on Academy Awards for best actors was gathered from Wikipedia.

Data Collection: Description of features

Features used in the model

Selecting and transforming independent variables is a pivotal step in developing highly accurate predictive models. In the context of analysing movie success, past research has explored three primary types of features: **audience-based** (e.g. comments from Twitter or movie reviews), **release-based** (e.g. number of theatres a movie is shown in or season of release) and **movie-based** (e.g. age rating, genre, cast, crew).

Our project focuses on predicting movie profitability during the pre-production stage, limiting our ability to utilize most audience and release-based features, as they aren't available when making investment decisions. In light of this constraint, we incorporated the following features into our models, with some of them extracted considering previous research literature:

- **Genre**: Categorical feature encompassing 17 genres (e.g. Action, Adventure, Drama, Comedy). Documentary movies (15 in our sample) were excluded from the data as these are typically not released to theatres and may not involve professional actors.
- **Age rating**: Categorical feature covering 6 ratings (e.g. PG, 15+).
- **Sequel category**: Binary feature indicating whether a movie belongs to a collection (e.g. Ice Age Collection), excluding the first movie in a collection given the first movie is not supposed to have an established fan base. It's important to note that due to the 20-year coverage limit, some movies within the dataset might have been misclassified as initial movies of collections, resulting in their exclusion from the Sequel category.
- **Runtime**: Duration of the movie.
- **Budget**: Estimate of production expenditure excluding other costs such as marketing or distribution.
- **Director’s metrics**: Features capturing the director's average ROI and average profit from previous movies, reflecting the director's track record.
- **Actor’s metrics**: Features indicating the highest average ROI and highest average profit from previous movies of one of the three main actors in the movie, highlighting the key actor's past success.
- **Number of Oscars**: Total Oscars won by the three main actors before the release of the movie.
- **Previous year's average worldwide annual box office**: Average box office across all movies in the year prior to the release of selected movie. This aims to capture the overall economic state of the film industry before the movie is released.
- **Annual Profitability Percentage by Genre**: Percentage of profitable movies with the same genre as the selected movie in the year before its planned release.

Preprocessing: Data Cleaning

- **Removing duplicates**: Duplicate records, if present, can skew analyses and model performance. We conducted an assessment to identify duplicates within the dataset. Fortunately, no duplicate entries were found, ensuring data integrity and accuracy.

- **Handling missing values:** Missing data can significantly affect analysis outcomes. We systematically addressed missing values by:

  - **Removing entries with budget or revenue below $5,000**: A total of 6,188 observations were removed due to insufficient budget or revenue data, as both budget and revenue are pivotal for calculating ROI, our target value.

  - **Eliminating entries with missing age ratings**: 201 observations were excluded from the dataset due to missing age rating values.

  - **Removing missing director or actor metrics**: Encountering a notable number of missing values (1,697 observations) for director and actor metrics due to a 20-year data limit prompted a critical decision. We compared two options: removing missing values causing substantial data loss or retaining observations while removing these features. Evaluating two models—one including all features but fewer observations and the other maintaining all observations but excluding these features—we observed that the model with all features but fewer observations performed significantly better in precision-recall ratios. Consequently, we chose to retain director and actor metrics, removing the observations with missing values. Although this reduced the dataset size, it notably enhanced the model's performance, aligning with our goal of optimising precision-recall ratios for predicting movie profitability."

- **Outlier detection and treatment:** Outliers are observations significantly different from other data points that can significantly impact the performance and accuracy of predictive models. In our data preprocessing phase, we employed the following methods to identify and manage outliers effectively.

1. Identification of outliers: Use of statistical methods (Z-score, IQR and Tukey's method) and visual tools (box plots, histograms, and scatter plots) through exploratory data analysis (EDA) to visualise numerical feature distributions. After evaluating the distribution of ROI values in the dataset, it was observed that the majority of movies fall within a reasonable range, with the 75th percentile at around 3 and the 90th percentile at approximately 6. However, 21 movies exhibited exceptionally high ROI values (greater than 40), significantly deviating from the dataset's general trend.

2. Treatment Methods for Extreme ROI Values - **Removal of movies with an ROI higher than 40 for the following reasons:**

   - Influence of Outliers: These values represent outliers not typical of industry movie behaviour, removing them aids in building a model that better represents dataset trends.

   - Enhanced Model Robustness: Removal of extreme outliers can lead to a more robust and reliable model, reducing overfitting and enhancing performance on unseen data.

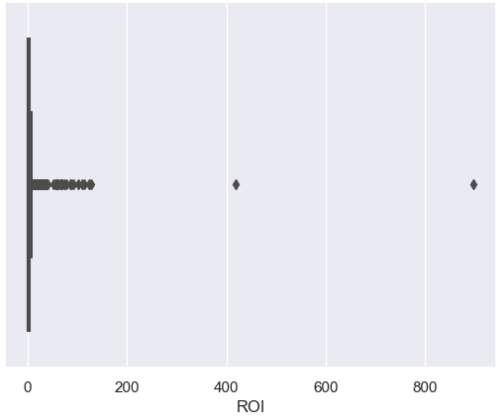
The following slide shows the impact of removing these outliers.

Preprocessing: Excluded movies with extreme ROI

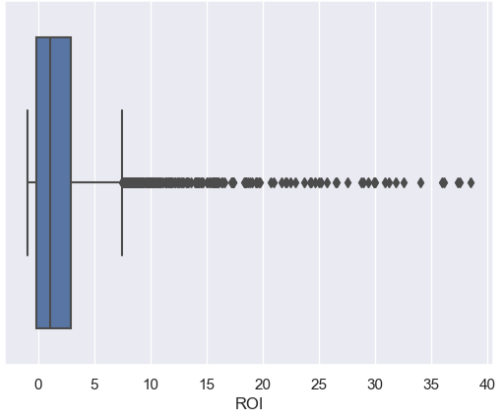
It's important to note that the movies with exceptionally high ROI values (greater than 40) that were excluded from the dataset might represent a **unique subset of highly successful movies**. Specifically, these movies share a common trait of having budgets below £5 million but attaining remarkable profitability. While these outliers were removed to ensure the robustness of the model and align the analysis with the general trends of the dataset, acknowledging their existence is crucial due to their distinct characteristics:

- **Uniqueness in success:** These movies, despite their relatively lower budget, achieved extraordinary ROI, indicating exceptional success compared to the average successful movie. Their exclusion from the dataset doesn't diminish their significance but rather recognises their atypical nature.
- **Representation of extreme success cases:** While **not representative of the typical or average successful movie**, these exclusions signify a separate category of movies that have defied the norm in terms of profitability within a constrained budget.
- **Potential separate evaluation:** Evaluating these excluded movies separately could provide valuable insights into the factors contributing to their exceptional success. This is not within the scope of our project but since there are only 21 movies with these characteristics, the following methodologies could be pursued: In-depth case studies, comparative analysis or getting industry expert insights.

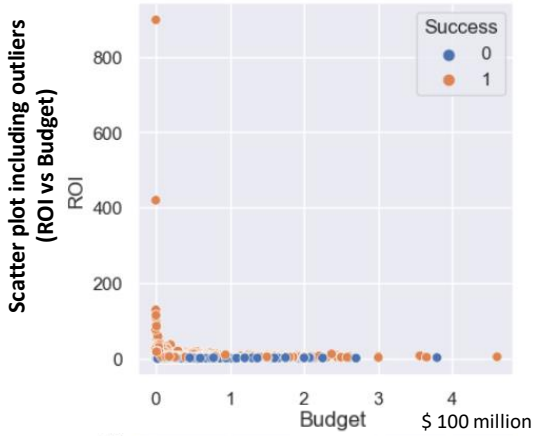
ROI box plot including outliers



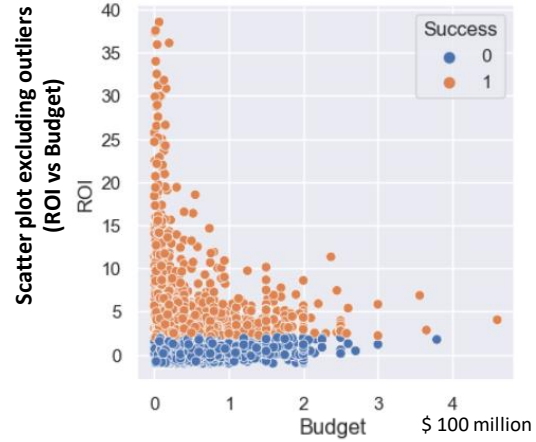
ROI box plot excluding outliers



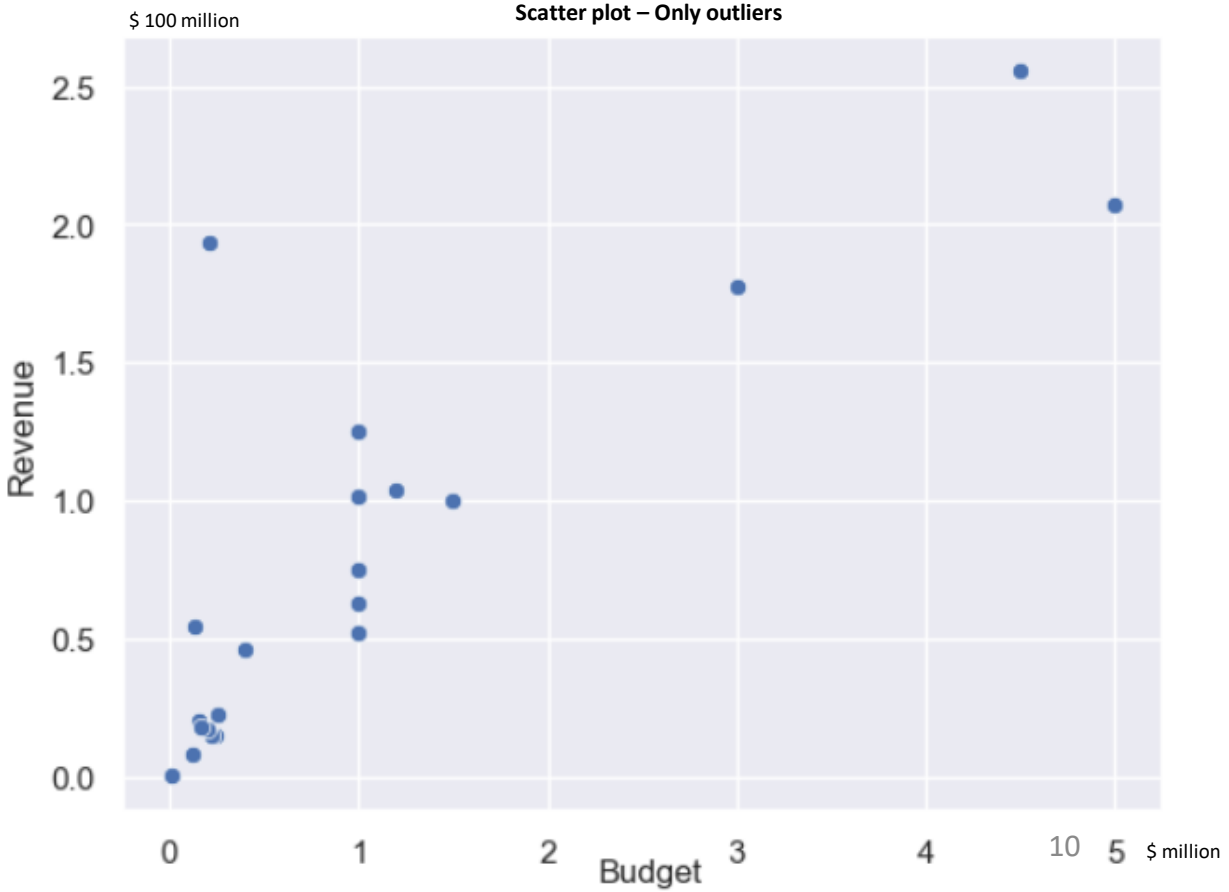
Scatter plot including outliers (ROI vs Budget)



Scatter plot excluding outliers (ROI vs Budget)



Scatter plot – Only outliers



Preprocessing: Feature Engineering and Data Transformation

Feature engineering and data transformation are pivotal steps in preparing the dataset for modelling. These processes involve creating new features, modifying existing ones, and standardizing the data to enhance the model's predictive power and performance.

Feature engineering

- **Genre and Age Rating Encoding:** Both features, genre and age rating, were encoded using one-hot encoding to convert them into numerical values suitable for modelling.
- **Sequel Identification:** The presence or absence of movies belonging to a sequel was transformed into a binary feature to capture the influence of movie collections on profitability.
- **Director and Actor Metrics:** Metrics related to directors' and top actors' historical success in previous movies, such as average ROI and profit, were calculated and incorporated as features to capture their influence on a movie's success.
- **Oscars and Movie Trends:** Features indicating the number of Oscars won by main actors and the profitability percentage of movies in the same genre as the selected movie in the previous year were engineered to reflect industry trends and acclaim.
- **Previous Year's Average Worldwide Box Office**: This feature was created to capture the average box office performance across all movies in the year before the release of a selected movie, providing insight into the overall economic state of the film industry before a movie's release.

As part of the **feature selection process**, we used Select From Model and Recursive Feature Elimination (RFE) approaches for those models where feature importance can be estimated directly (Logistic Regression, Random Forest and Gradient Boosting).

Data transformation

- **Standardisation:** Numerical features were standardised using z-score normalisation to ensure all variables were on the same scale, preventing any particular feature from dominating the model due to differences in magnitude.
- **Target Variable Transformation:** Considering the binary nature of the target variable (positive ROI above or equal to 2 and negative ROI below 2), it was transformed to facilitate model training and evaluation, ensuring it aligns with the intended classification task.

Model validation

To assess the generalisation capability of our models, we employed data splitting techniques such as train-test splits or k-fold cross-validation (with 5 folds). This process involved partitioning the dataset into training and testing subsets. The training set was utilised for model training, using cross validation to define the best parameters for different evaluation metrics (accuracy, precision, F1, ROC AUC and average precision), while the test set allowed for unbiased evaluation of model performance on unseen data.

Conclusion

The preprocessing stage focused on enhancing the dataset's quality, ensuring it is suitable for training robust predictive models. Feature engineering enabled the creation of informative variables, while data transformation techniques standardised and prepared the data for modelling. The process aimed to optimise predictive accuracy while maintaining the integrity and representativeness of the dataset.

Exploratory Data Analysis (EDA)

EDA: Introduction

Exploratory Data Analysis (EDA) is a critical preliminary step in understanding our dataset. It allows us to uncover patterns, anomalies, and relationships within the data, providing essential insights for subsequent analyses and modelling.

Overview of Dataset

Dataset Size: EDA based on 3,187 movies but the number of movies was reduced to 1,490 movies after removing the missing values for director and actor metrics. We initially started with 49 columns and we ended up selecting 13 columns (including the target) to develop the model. These 13 attributes have been described in the Data Collection section of the report and will be further explored in the EDA section.

The dataset used for our EDA has 3,682 actors, 1,621 directors and 1,408 movie studios. We have shown some visualisations of these important film contributors.

Objectives of EDA

- For categorical variables: Show frequency distributions, bar charts, or pie charts.
- For numerical variables: Display summary statistics (mean, median, range, variance), histograms, or box plots.
- Investigate relationships and dependencies among variables using techniques such as pair plots and correlation matrixes.
- Identify potential outliers and anomalies (this has been already covered in the Preprocessing section of the report).
- Gain preliminary insights for feature engineering and model preparation.

Let's delve into our dataset to reveal the fascinating insights lying within!

Movie trends in our data

1. Number of movies per year

The bar chart illustrates the annual distribution of movies within our dataset, revealing a consistent range of 100 to 200 movies released each year. Notably, we observe an increase in number of movies during the first decade, followed by a notable decline in 2020 attributed to the COVID-19 pandemic's impact on the film industry.

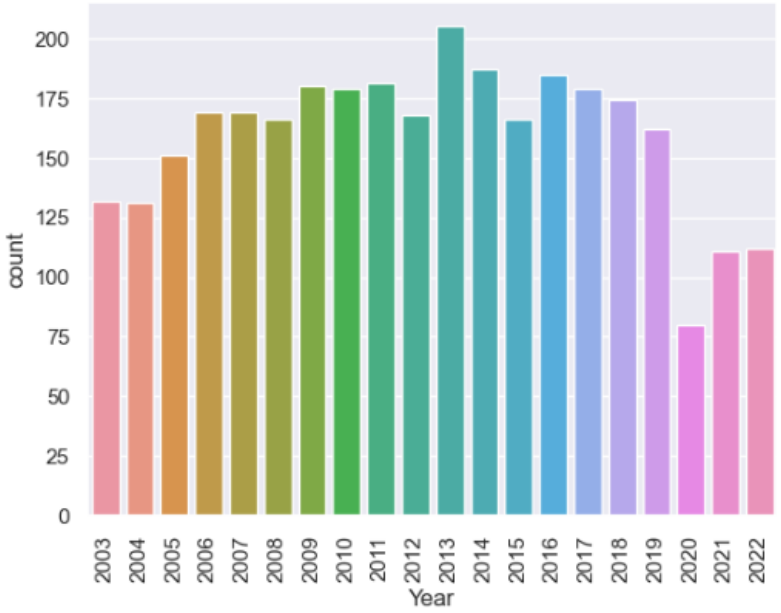
2. Average budget and revenue trends over years

The line chart shows the average budget and revenue trends for the movies within our dataset across different years. The average budget has remained relatively stable, hovering around $50 million. Regarding average revenue, we can observe a steady increase from c.$100 million in 2006 to c.$170 million in 2019. However, a sharp decline to around $50 million occurred in 2020, followed by a robust recovery to approximately $150 million by 2022.

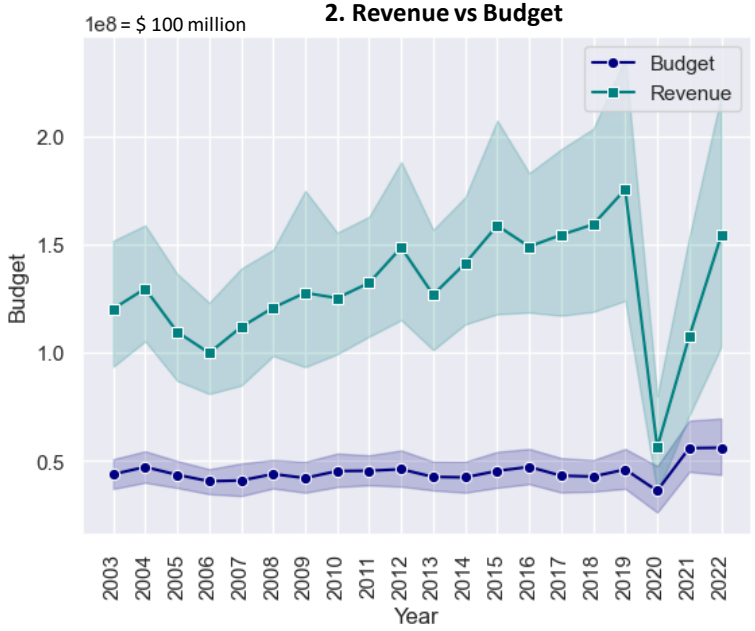
3. Percentage of successful and unsuccessful movies per year

This bar chart the percentage distribution of successful and unsuccessful movies yearly, based on our ROI threshold for defining success (ROI equal or greater than 2). It provides insights into performance trends of movies within our dataset annually. Notably, the data shows that the average success rates fluctuate between 30% and 40% with the exception of 2020 and 2021, when it declined to approximately 20% due to impact of the pandemic around the globe.

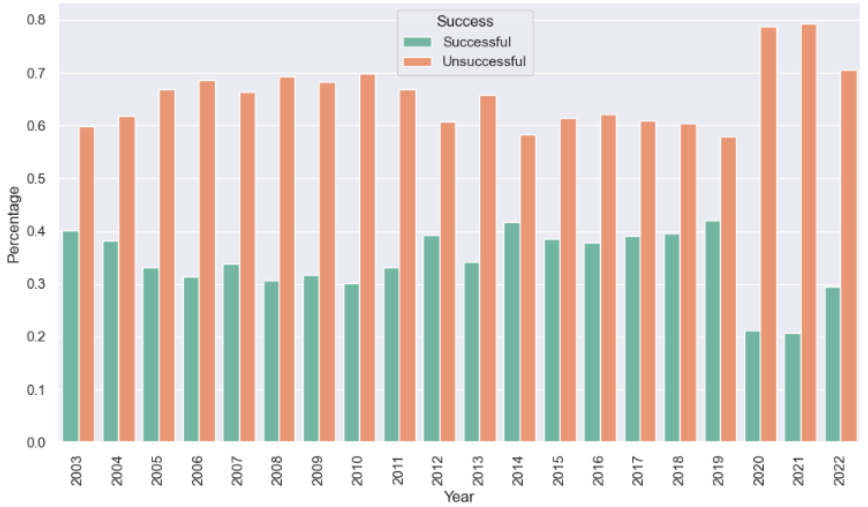
1. Number of movies



2. Revenue vs Budget



3. Success rates per year

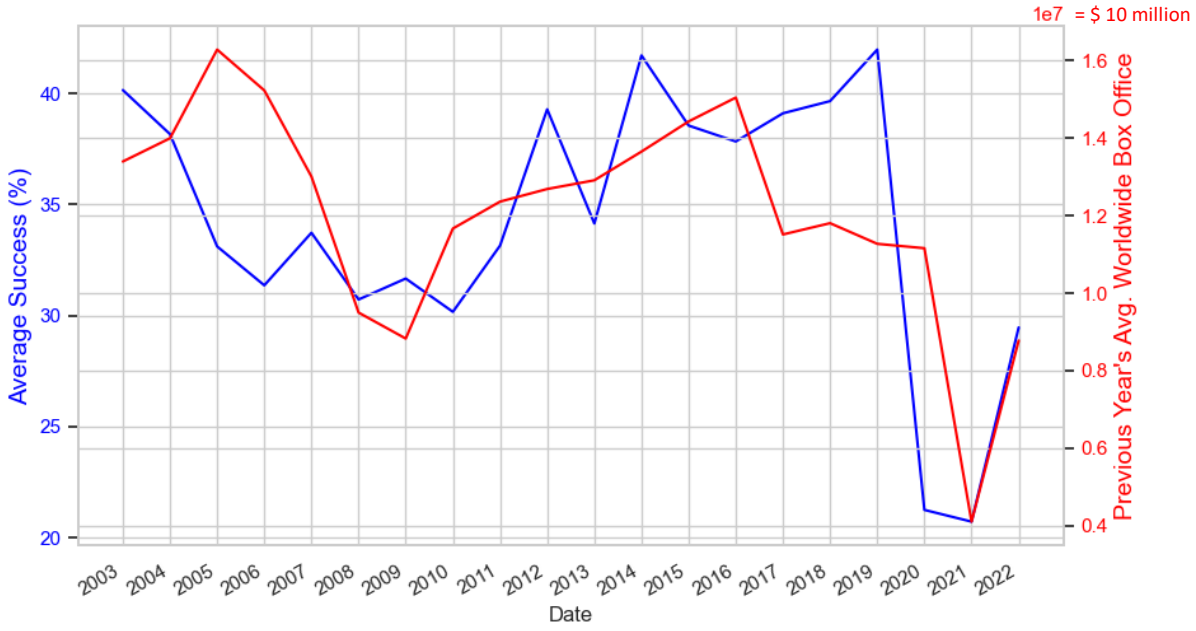


Impact of pre-release industry dynamics

The average worldwide box office performance across the film industry (extracted from BoxOfficeMojo) from the preceding year (relative to individual movie release) has been incorporated into our predictive models. This inclusion aims to capture the film industry's overall economic state before a movie's release.

Movies are often influenced by the industry's previous year's performance, which significantly impacts their commercial success. A robust average worldwide box office for the preceding year could indicate increased audience interest, a strong market, and heightened anticipation within the industry. This, in turn, could also positively correlate with the potential success of a movie, suggesting a favourable environment for audience reception and financial returns.

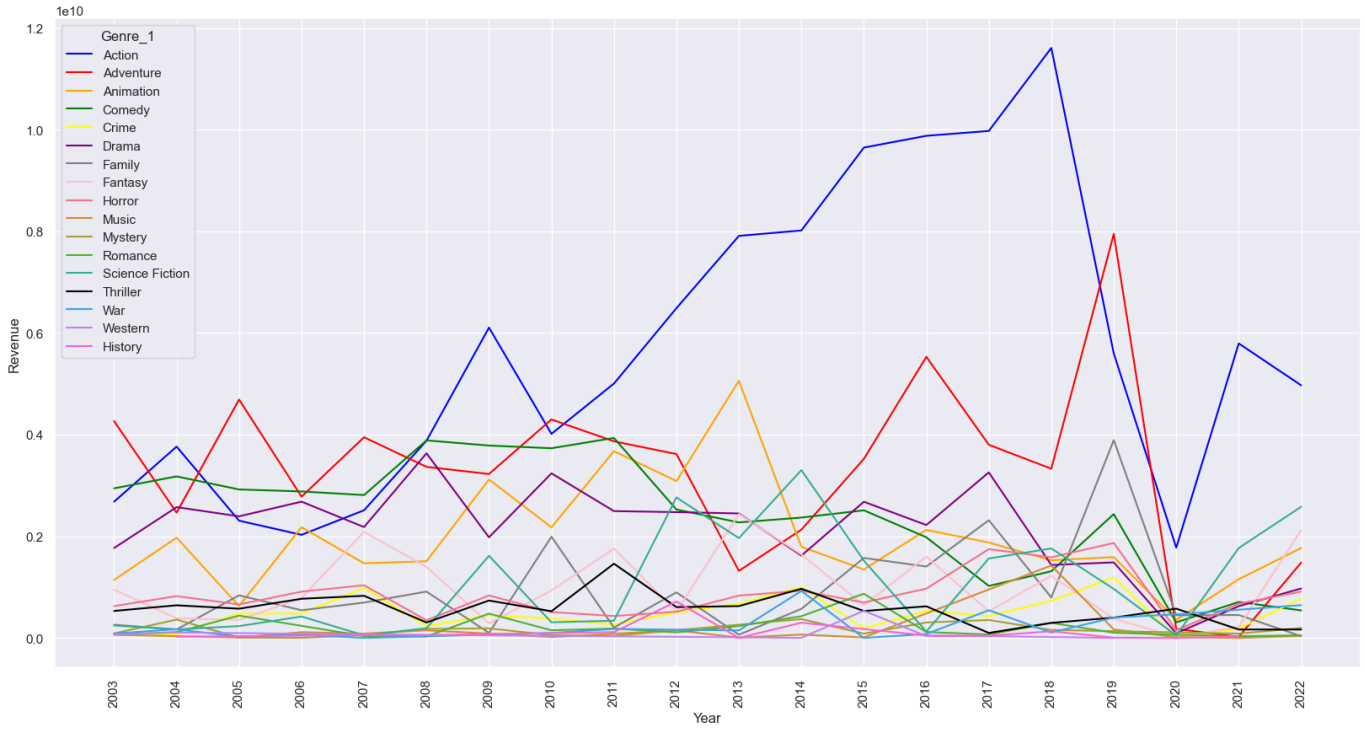
The chart below aims to show the impact of industry's overall economic state, measured using Previous Year's Average Worldwide Box Office (PYAWBO), on average success rate of movies within our dataset. The horizontal axis illustrates the progression of time, showcasing the years considered in our dataset. Featuring two vertical axes, the left axis displays the average success rate of movies (ROI greater than 2) that are part of our dataset in each year while the right axis shows the Previous Year's Average Worldwide Box Office.

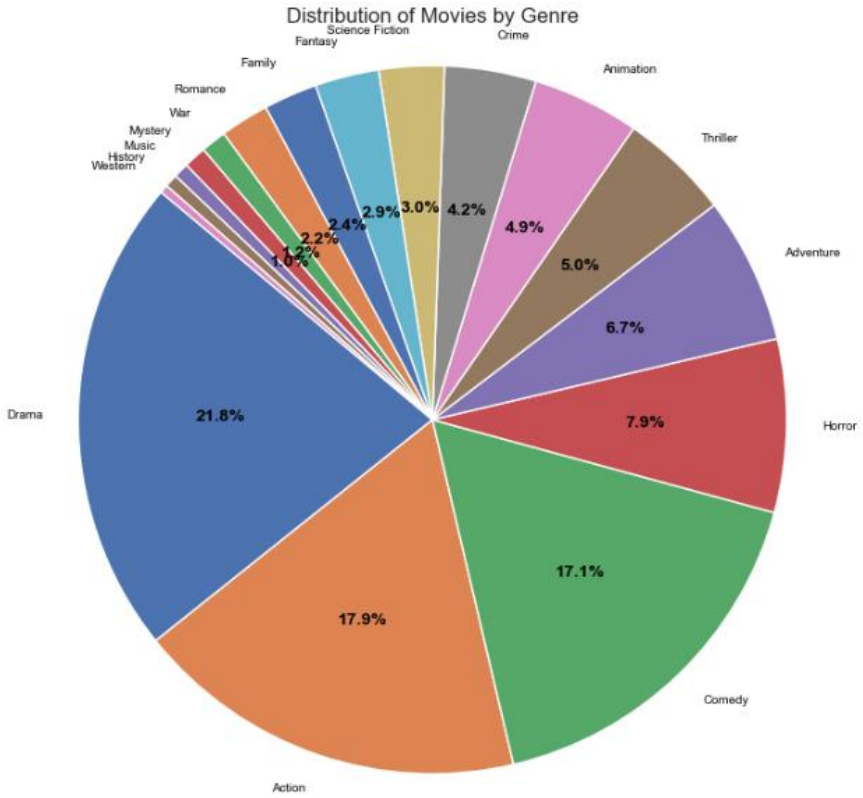


Our analysis reveals a moderately positive correlation of 0.57 between the industry's global box office performance and the average success rate within our dataset in the subsequent year. Surprisingly, despite the observed relationship, the correlations of PYAWBO with both movie ROI and movie success are negligible (0.06), suggesting a divergence between aggregate and individual impacts.

In conclusion, the 0.57 correlation between PYAWBO and average success rate might indicate a relationship at an aggregate level, reflecting broader industry trends. However, this correlation does not hold true for individual movies due to various unique factors that influence their success or failure. Understanding movie success necessitates consideration of a broader spectrum of variables beyond global box office performance.

Genre distribution and revenue trends





1. Movie genre distribution

The pie chart displays the distribution of movie genres within our dataset. The most predominant movie genres are Drama (c.22%), Action (c.18%) and Comedy (c.17%), collectively accounting for approximately 57% of the total movies in our sample.

2. Revenue trends by genre

The line plot shows the revenue trends across different movie genres from 2003 to 2022. Notably, Action and Adventure genres consistently drove the highest total revenue over the years. Additionally, Comedy, Drama, and Animation genres have shown substantial success in generating revenue. Family and Science Fiction genres exhibited periods of relatively high revenue within specific time frames.

Financial performance by Genre

1 & 2) Budget and Profit box plot by Genre

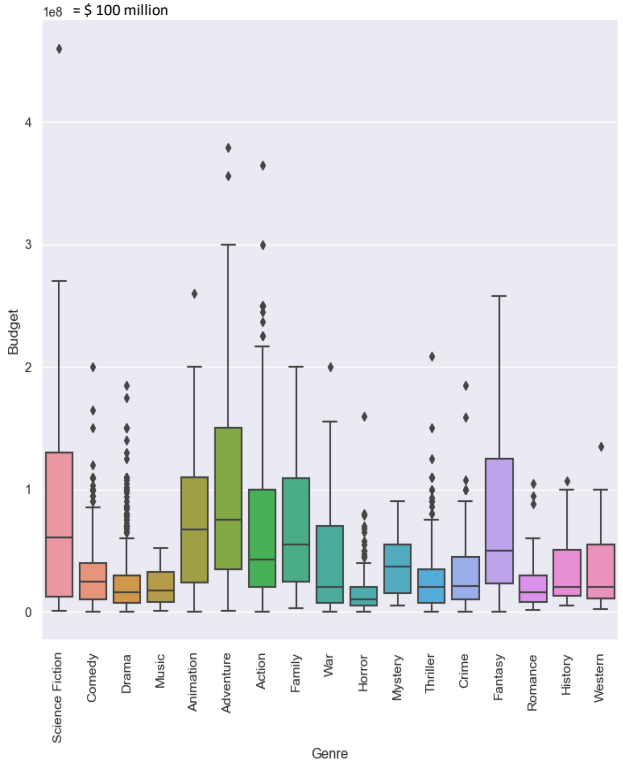
Displayed in a box plot format, this visualisation identifies budget and profit ranges, medians, and outliers. We limited the profit box plot's "y axis" to make sure it would have the same scale ($100 million) as its budget counterpart, which is helpful for comparative purposes but hides some of the outliers. The genres with largest profits such as Animation, Adventure and Fantasy, are also the ones with largest budgets. Action and Adventure movies have the largest profit outliers (not shown in the chart), with both of them having around 7 movies with profits above $1,000 million.

3) Percentage of successful and unsuccessful movies by Genre

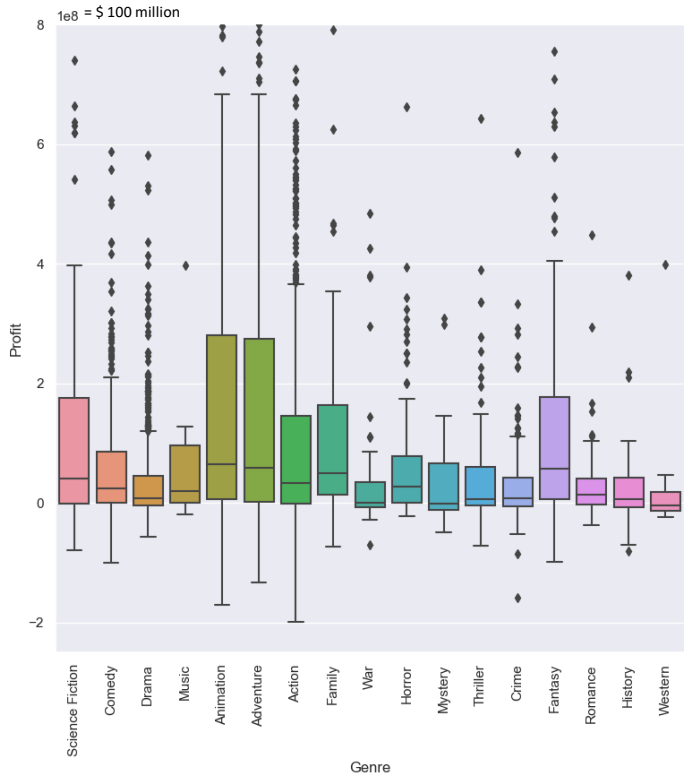
The bar plot presents the percentage of successful and unsuccessful movies categorised by genre, offering an overview of which genres tend to achieve profitability (ROI equal or greater than 2) and those that may struggle in this aspect. Most genres had a successful rate between 30% and 40%, which is aligned with what we previously observed when we looked at the success rates by year. War, Western and Crime had successful rates below 30% whereas Music and Horror were the most successful genres, with a success rate around 50%.

Upon examining the three charts**, it's apparent that there isn't a direct correlation between a genre's profitability and a high ROI** mainly because highly profitable films within certain genres were associated with very large budgets. This observation suggests that while certain genres, such as Music and Horror, demonstrated higher success rates around 50%, this may be attributed to being among the genres with lowest budgets rather than solely the genre itself.

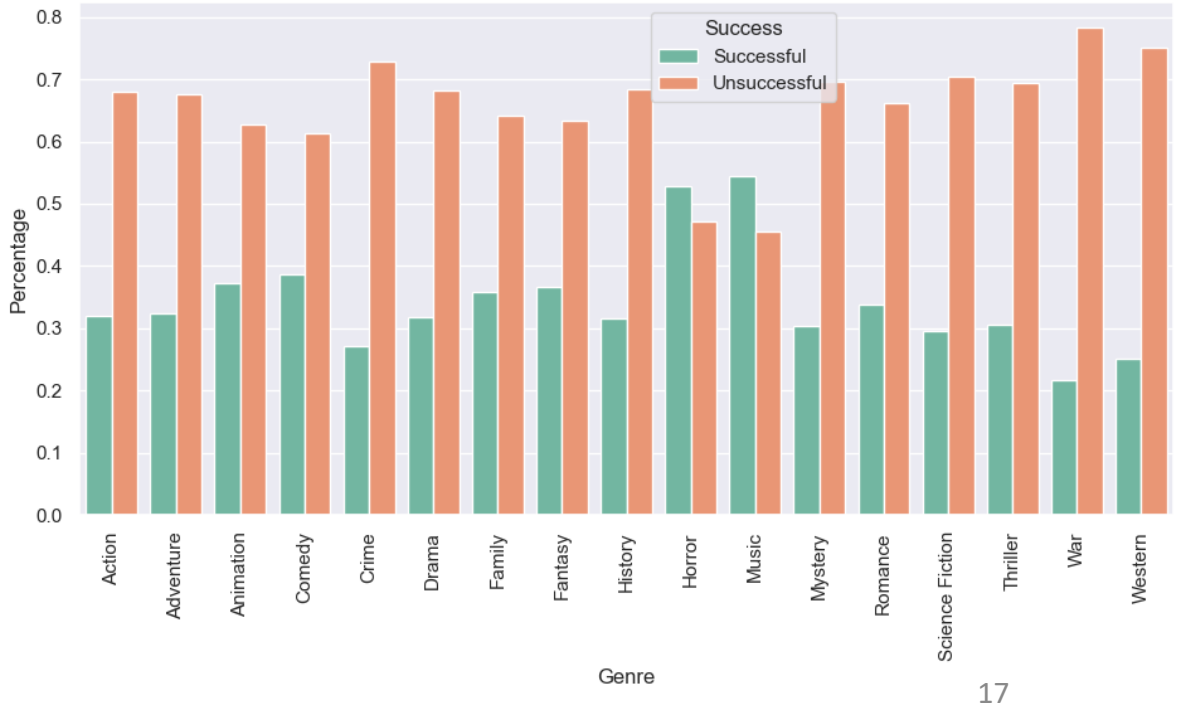
1. Budget box plot by Genre



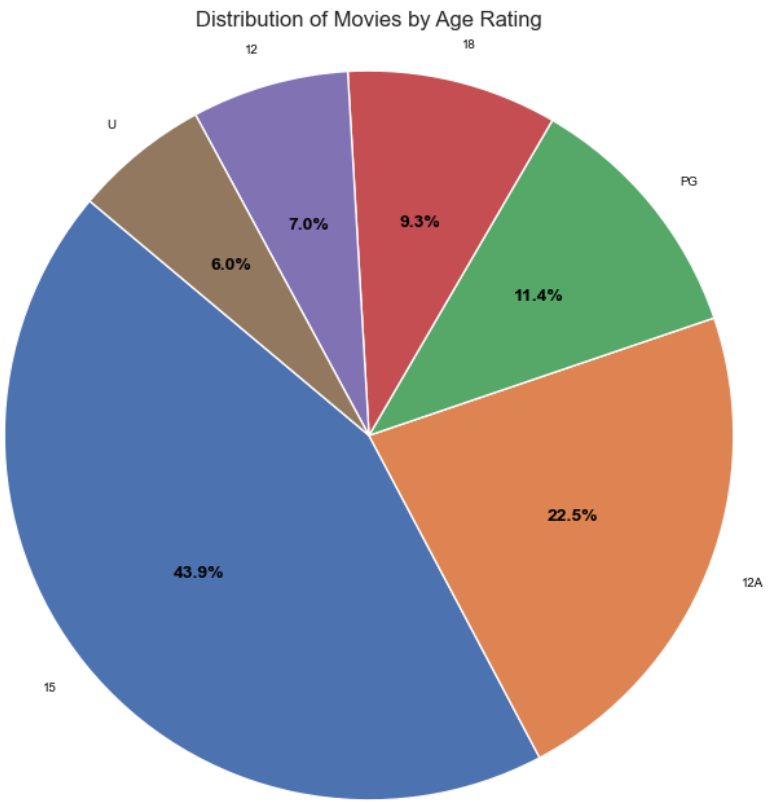
2. Profit box plot by Genre



3. Success rates by Genre



Age rating distribution



A movie age rating, also known as content classification, is a categorisation assigned to a film by a regulatory body or institution based on its content. These ratings offer guidance to audiences regarding a movie's suitability for different age groups. Since our project is based on movies released in the UK, the age ratings are the following:

- U (Universal): Suitable for all
- PG (Parental Guidance): For general viewing but some scenes may be unsuitable for young children
- 12A: Generally suitable for viewers aged 12 and above, however children under 12 can watch the film in a cinema if they are accompanied by an adult
- 12: Suitable only for individuals aged 12 and above
- 15: Suitable only for 15 years and over
- 18: Only for adults

Other countries have different age ratings but we would expect that movies released globally, which represent almost all our data, would have equivalent age ratings in other regions.

Considering movie age ratings is important to predict financial success for several reasons:

Market Appeal and Viability: Movies aimed at specific age groups tend to have distinct themes, genres, or content that resonate more with those audiences. The target age rating can influence a film's market appeal and its potential success at the box office.

Box Office Performance: Different age ratings can significantly impact the size of the movie's potential audience. Films rated for a broader audience (e.g., U or PG) might attract larger crowds, potentially leading to higher box office earnings. Conversely, movies with more restrictive ratings might limit their audience but could still generate substantial revenue through targeted marketing efforts.

Industry Trends and Success Factors: Analysing the financial success of movies across different age ratings can reveal industry trends, aiding in predicting and understanding how various types of movies perform financially.

In summary, considering age ratings for movies provides valuable insights into the target audience, marketability, content restrictions, and potential box office performance. Incorporating age ratings into predictive models for movie profitability offers a more comprehensive understanding of the factors influencing financial success.

Financial performance by Age Rating

1 & 2) Budget and Profit box plot by Age Rating

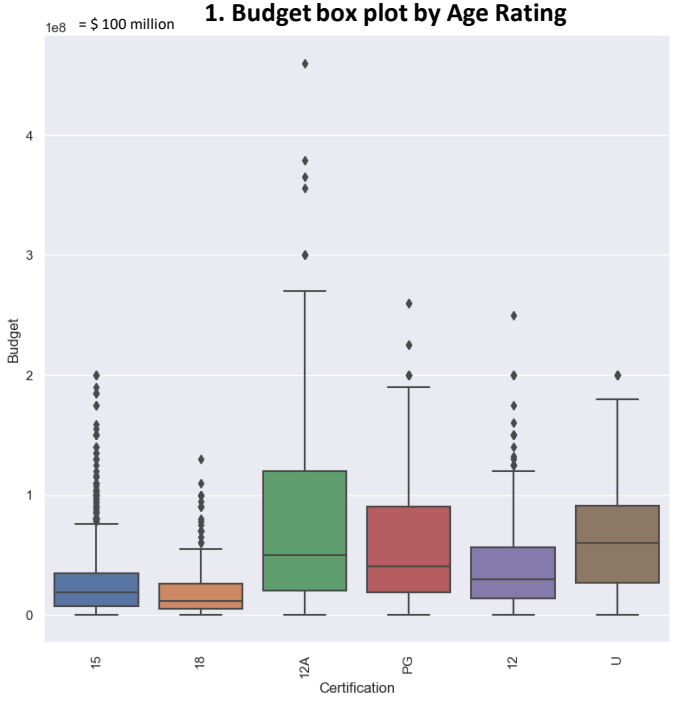
The box plots illustrate the distribution of budgets and profits across various age ratings. We adjusted the profit box plot's scale to align with the budget plot, aiding in comparative analysis but concealing some outliers. Notably, age ratings such as U, 12A, and PG, which command the largest budgets, also demonstrate higher profitability. The 12A rating exhibits noteworthy profit outliers, approximately 15 movies reporting profits surpassing $1,000 million.

3) Percentage of successful and unsuccessful movies by Age Rating

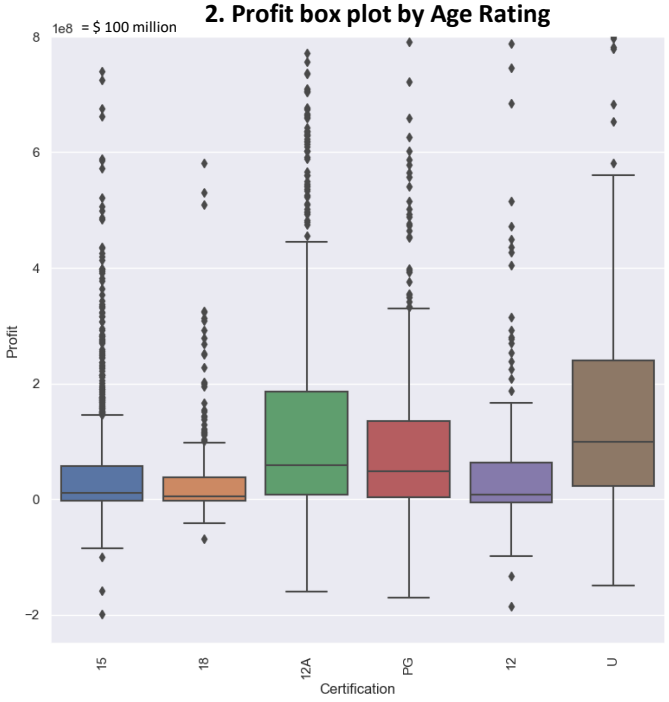
Across most age ratings, the success rates hover between 30% and 40%, consistent with earlier observations by year and genre. However, the 12 Age Rating reports success rates below 30%, while the U Age Rating emerges as the most successful, boasting a success rate between 45% and 50%.

Although average success rates vary among age ratings, it's crucial to consider that movies belonging to each rating category are crafted with unique themes, genres, or content intended to resonate more profoundly with their respective viewers and hence the differences we observe could also be explained by these factors.

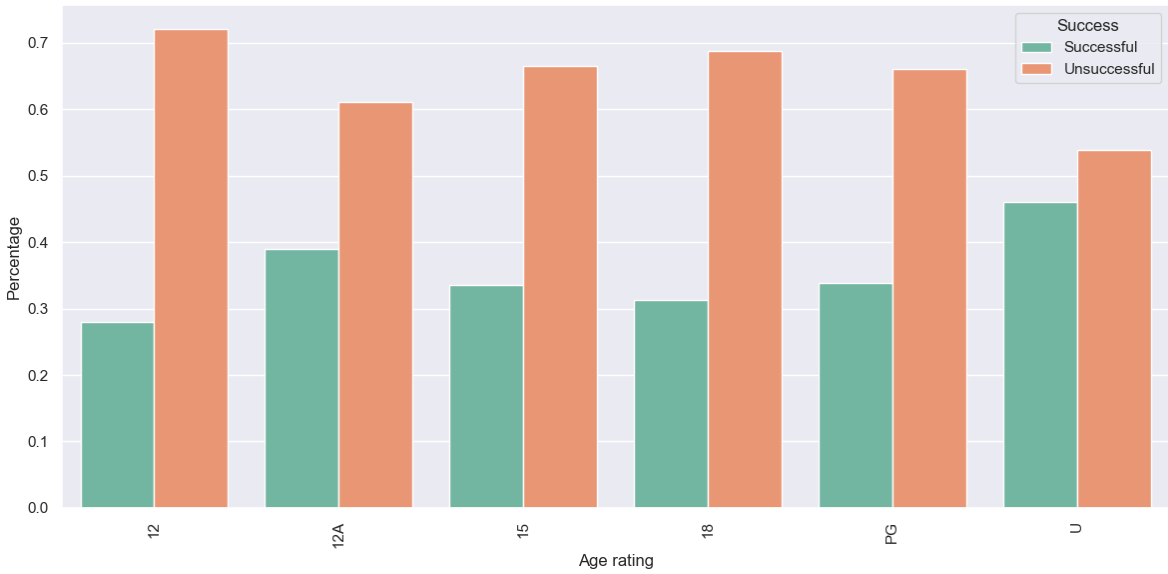
1. Budget box plot by Age Rating



2. Profit box plot by Age Rating



3. Success rates by Age Rating

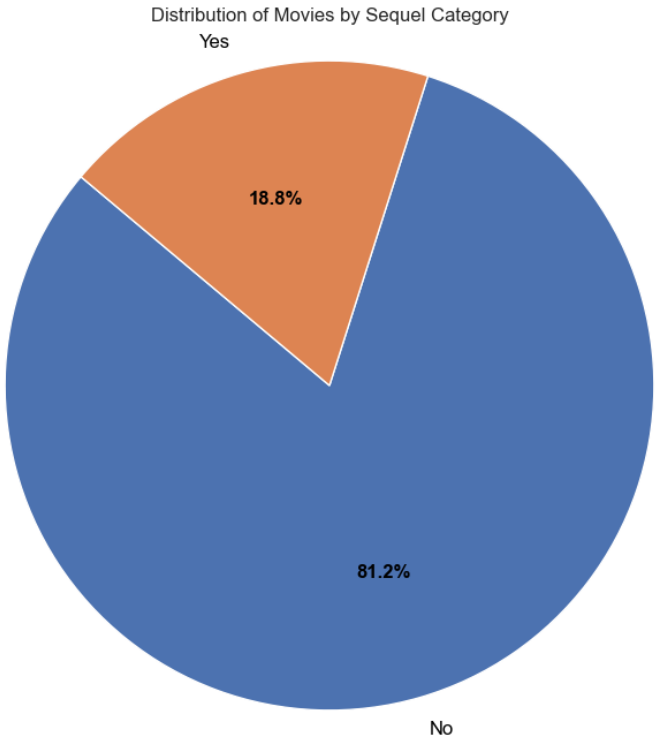


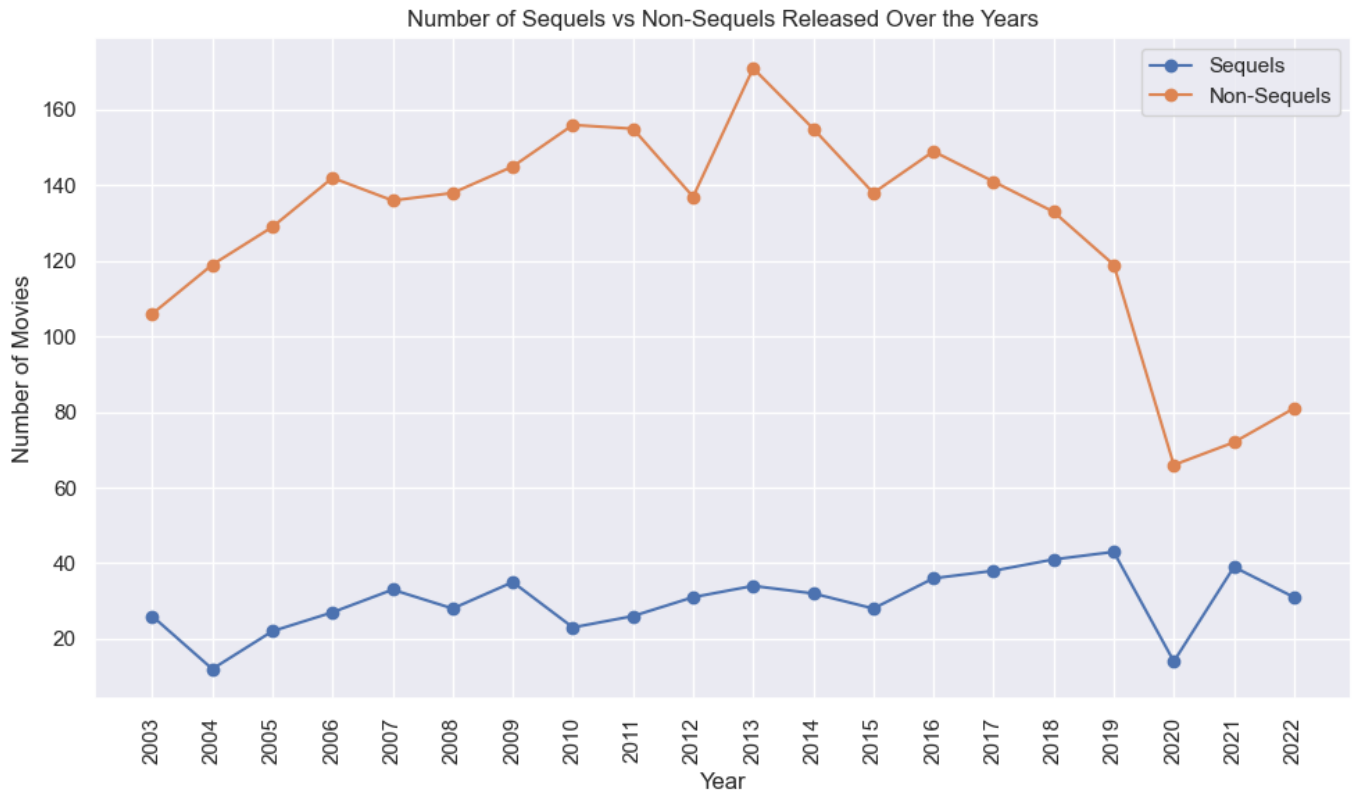
Sequel distribution and trends

A sequel refers to a subsequent film that continues or extends the storyline of a previously released movie. It's a cinematic follow-up that typically features recurring characters, themes, or elements from its predecessor and often aims to capitalise on the success and familiarity of the original movie. Considering the presence of a sequel in movie prediction models holds significant importance for several reasons:

1. **Audience Engagement and Expectations:** Sequels often attract an audience familiar with the original film, leveraging their established connection to the storyline or characters. This built-in audience anticipation can significantly influence box office performance.

2. **Risk Assessment and Investment Strategy:** The presence of a sequel in a movie dataset serves as a proxy for a film's potential to garner attention and generate revenue. Investors and production studios can use this information to assess risk and make informed decisions regarding resource allocation and investment in movie production.





Financial performance for Sequels

1 & 2) Budget and Profit box plot

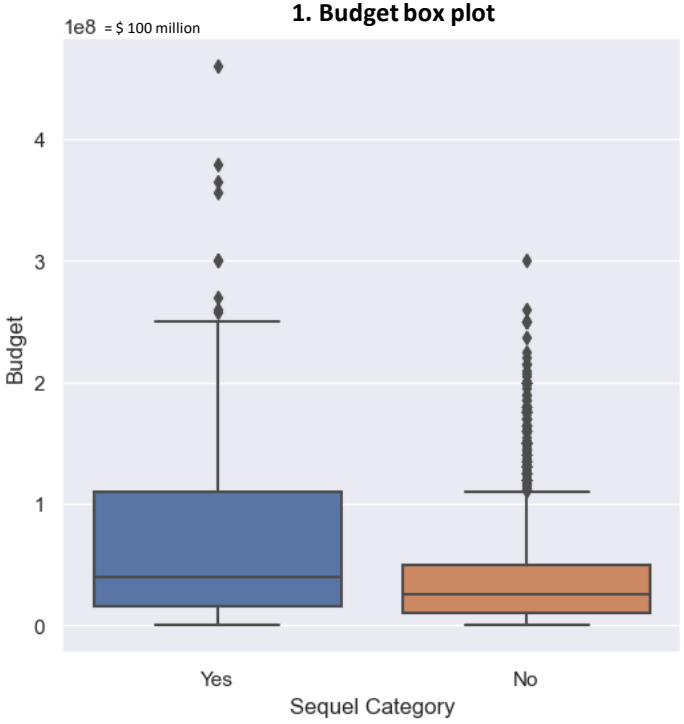
The box plots illustrate the distribution of budgets and profits based on whether a movie is a Sequel or a standalone movie at the time of release. We adjusted the profit box plot's scale to align with the budget plot, aiding in comparative analysis but concealing some outliers. Movies that were part of a collection had substantially larger budgets and profits than the ones that weren't, denoting that studios increase their investments in a significant way when they see the opportunity to capitalise on movies that have an established fanbase.

3) Percentage of successful and unsuccessful

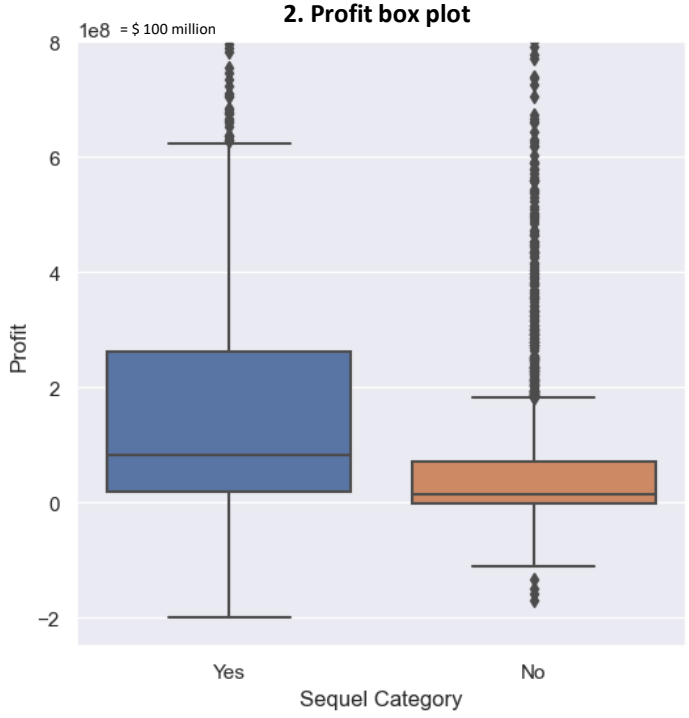
The difference in average success rate between movies that are part of a collection and those that are not is very significant. While only 30% of the standalone movies were successful (ROI equal or higher than 2), the success rate for the sequels was around 55%.

As we will see in more detail in the following sections of the report, **being part of a collection or franchise can be very financially rewarding**. This is probably due to sequels having an established fan base, brand recognition, more robust marketing campaigns among other factors.

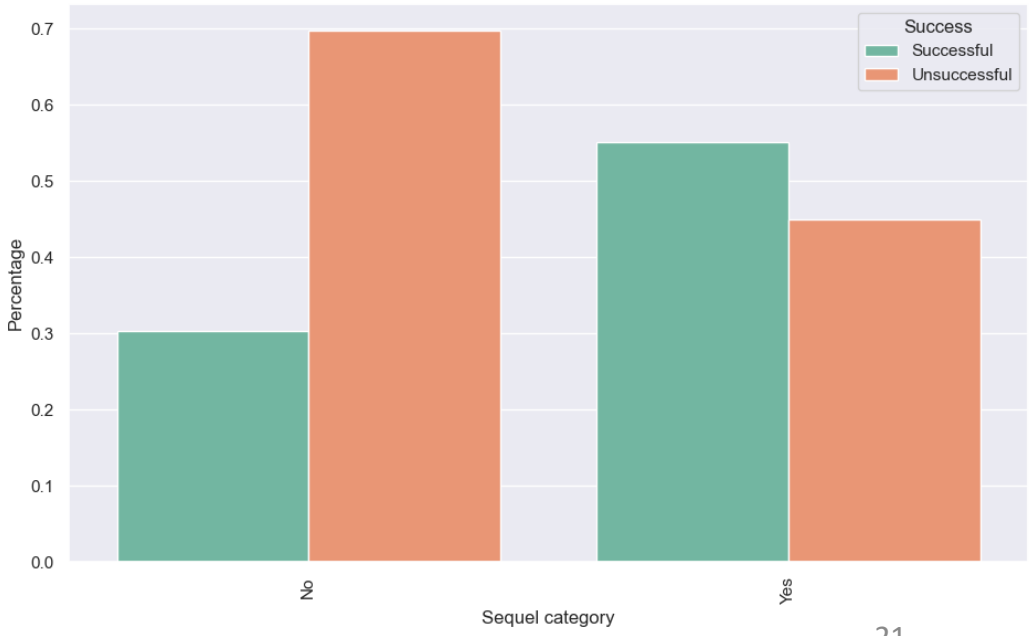
1. Budget box plot



2. Profit box plot



3. Success rates



Financial metrics

The pair plot presents the relationships between key financial metrics —budget, revenue, profit, and return on investment (ROI)— across the dataset, differentiating between successful and unsuccessful movies. It offers a comprehensive overview of the interplay among these variables.

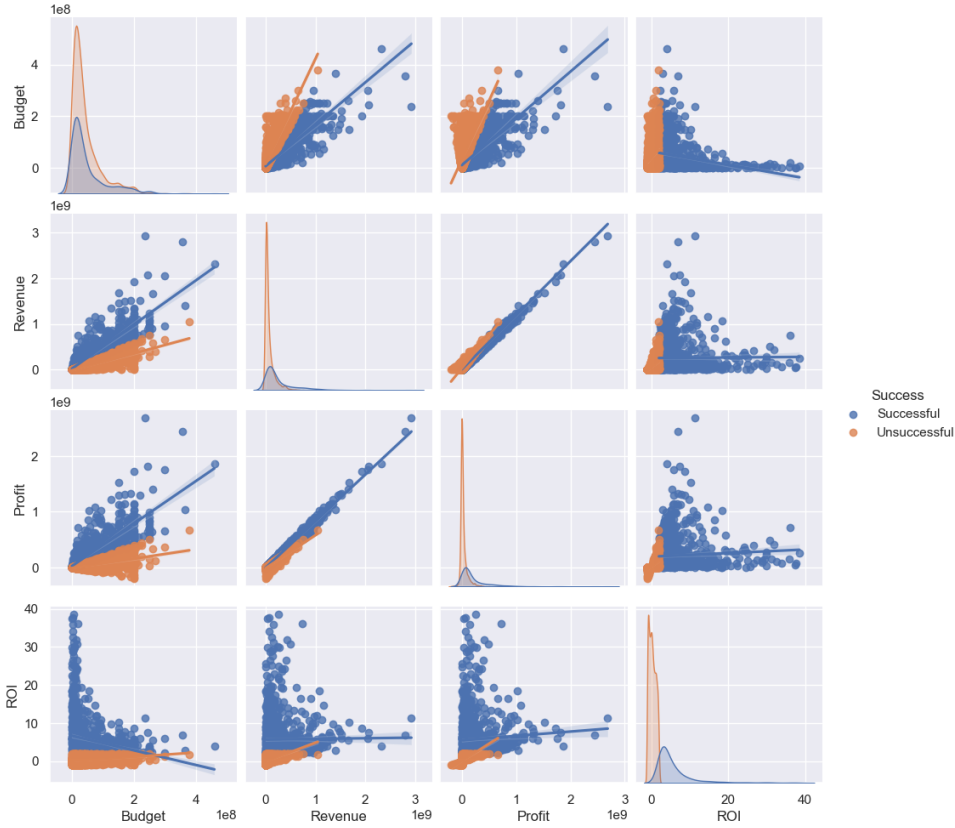
Notably, the histograms in the pair plot illustrate a significant disparity in the frequency distribution between successful and unsuccessful movies, primarily due to the **imbalance in the dataset**. **Successful movies constitute approximately 35% of the dataset**, leading to a higher count in the histograms for unsuccessful movies compared to their successful counterparts.

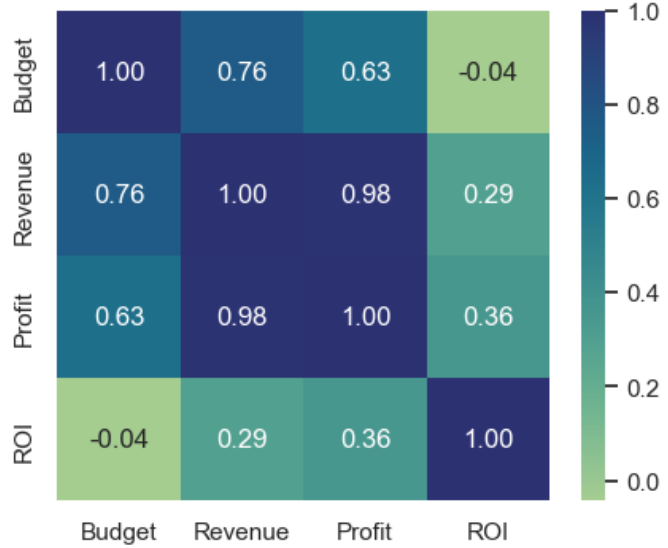
Additionally, a correlation heatmap was constructed to further explore the relationships between these metrics. The heatmap illustrates the degree and direction of linear relationships between pairs of variables, highlighting potential correlations among budget, revenue, profit, and ROI.

Key Observations and Interpretations:

- **Budget relationships**: Budget shows strong positive correlations with revenue and profit. For instance, higher production budgets tend to lead to larger marketing and distribution costs, positively impacting revenue and profit. However, there's an absence of correlation between budget and ROI, indicating that a higher budget doesn't necessarily guarantee a higher ROI due to increased challenges in achieving profitability.
- **Revenue, Profit and ROI**: Revenue and profit exhibit a strong linear relationship, as anticipated. Yet, both have moderately low correlations with ROI. This is because movies with larger revenue or profit often have larger budgets, leading to complexities in achieving a higher ROI despite increased earnings.

Conclusion: Successful movies tend to yield larger absolute returns (revenue or profit) with higher production costs. Nevertheless, achieving a superior relative return on investment (ROI) often resides in the realm of cost-effective crowd-pleasers, where each dollar invested generates a more substantial impact





Financial performance visualisation: Budget, Profit, and ROI dynamics

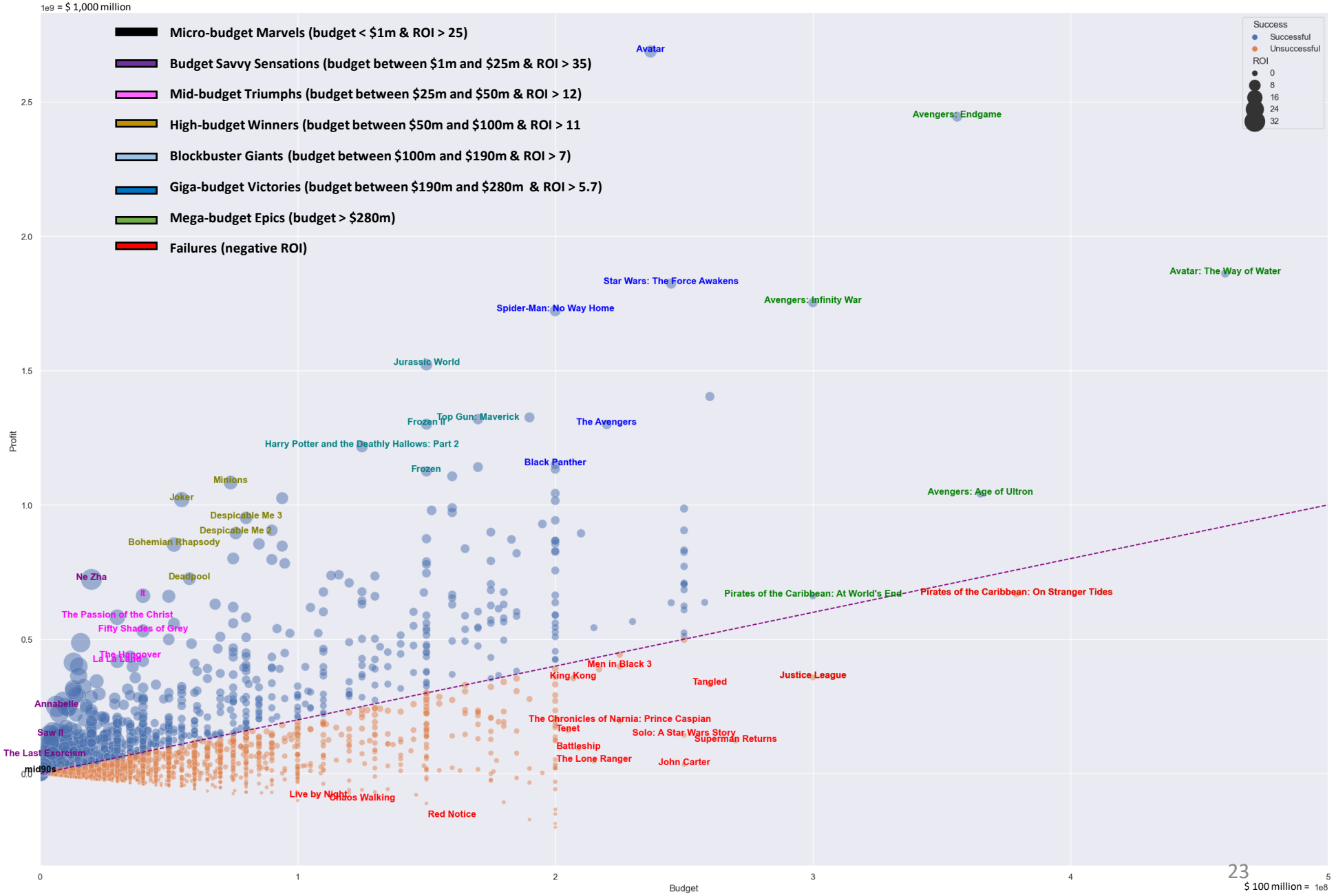
This visual representation shows approximately 3,200 movies released in Great Britain between 2003 and 2022.

Each bubble in the chart represents an individual movie, with its position along the horizontal (x) and vertical (y) axes denoting the budget and profit, respectively. The size and colour of each bubble correspond to the movie's ROI and success category. The text colours have been selected to highlight high ROI movies within various budget groups, ensuring clear differentiation without overlapping.

Budget vs. Profit: The placement of bubbles along the x-axis and y-axis showcases the distribution of movies concerning their budget and profit. Movies located toward the upper-right quadrant denote higher budgets and higher profits, while those positioned toward the lower-left quadrant indicate lower budgets and profits.

ROI as Bubble Size: The size of each bubble serves as an indicator of the ROI achieved by the movies. Larger bubbles represent movies with higher ROIs relative to their budget and profit, signifying their success in generating substantial returns despite the investment.

This visualisation aids in identifying movies that achieved remarkable financial success compared to their budget and profit, providing an intuitive representation to discern and patterns among movies in the dataset.



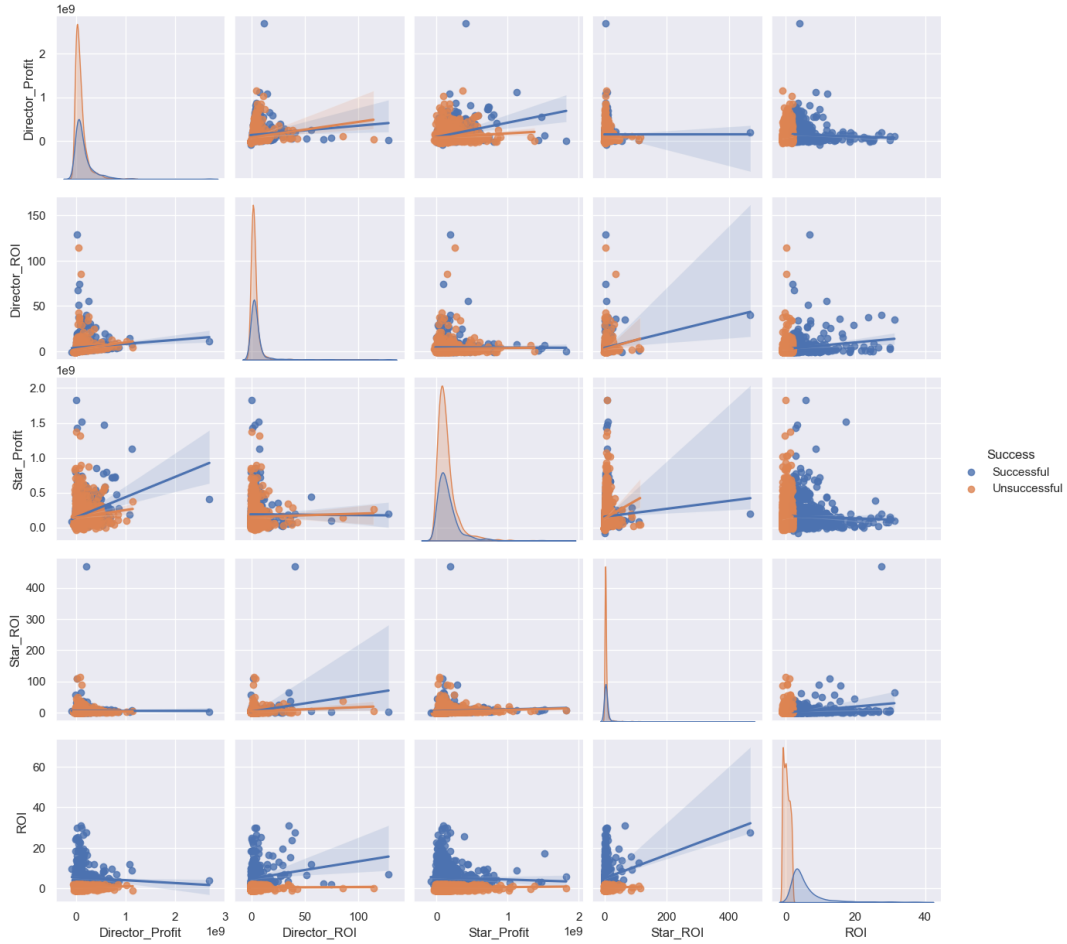
Director and Actor powers

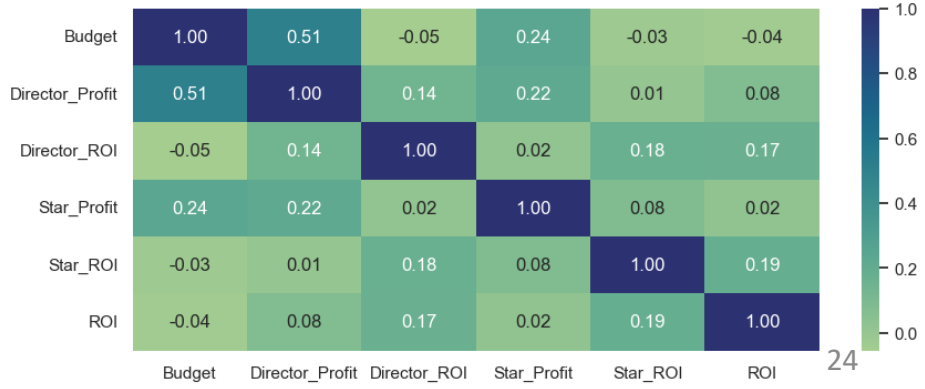
The pair plot visualizes the relationships and distributions between key performance metrics such as Director Profit, Director ROI, Star Profit, Star ROI, and ROI. These metrics offer insights into the historical performance of directors and key actors, as well as the overall return on investment for movies.

Additionally, a correlation matrix has been generated to explore the relationships among these metrics and budget. The matrix aims to uncover potential correlations and dependencies between the financial metrics and the historical performance indicators of directors and key actors.

Key Observations and Interpretations:

- **Director's previous average profit**: Exhibits a moderate positive correlation of 0.5 with the allocated movie budget. This moderate correlation might stem from various factors. Firstly, experienced directors with a history of delivering profitable movies might garner trust from studios, leading to larger budget allocations. Secondly, positive audience reception and a reputation for delivering largely profitable films could prompt investors to capitalise on a director's track record. However, the director's profit track record has a negligible correlation with the resulting ROI, indicating that a higher budget might not inherently guarantee a higher return.
- **Director's previous average ROI**: Demonstrates a low positive correlation with the movie budget (0.14) but exhibits a slightly higher correlation with the resulting ROI (0.17) compared to the correlation between the director's profit track record and ROI. This suggests that while a director's track record for delivering a high ROI might not influence budget allocation, it might have a slightly more pronounced impact on the resulting ROI of the movie.
- **Star Profit**: Shows a moderately low positive correlation with both the movie budget (0.24) and the director's profit track record (0.22). This alignment might echo the same reasons as discussed regarding the correlation between the director's previous average profit and the movie budget.
- **Star ROI**: Displays low positive correlations with Director's ROI track record (0.18) and movie ROI (0.19).





Director analysis: Budget, Profit, and ROI insights

This visual analysis encompasses approximately 360 directors responsible for directing multiple movies released in Great Britain from 2003 to 2022. To ensure clarity, directors overseeing at least 2 movies were included, preventing chart overcrowding while offering a representative sample.

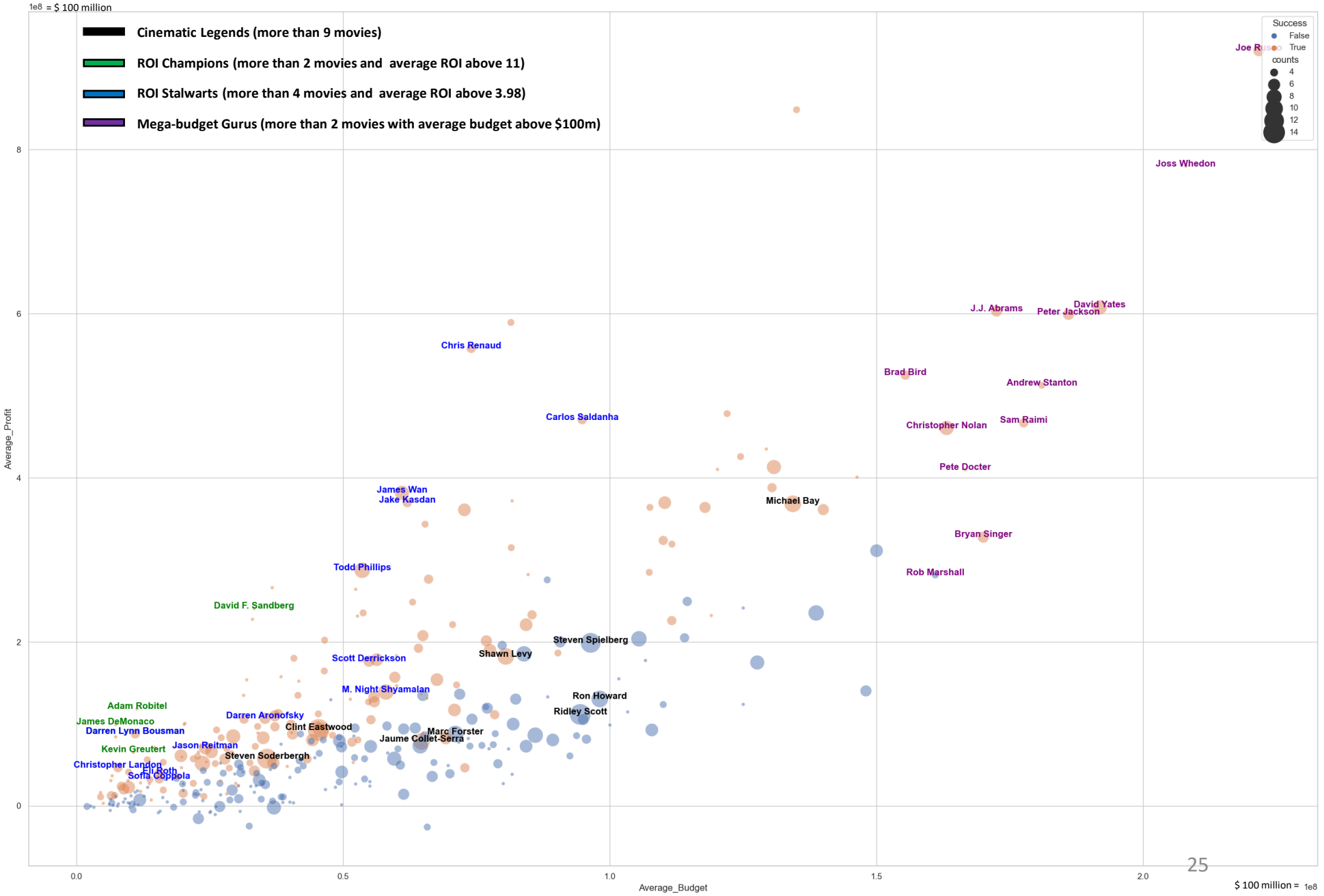
Each bubble on the chart represents a movie director. Positioning along the axes represents the average budget and average profit across their directed movies over the past two decades. Bubble size indicates the number of movies directed, while colour distinguishes directors based on whether their average ROI exceeds 2 or not. Text colours differentiate directors with diverse track records, ensuring clear visibility and differentiation.

It's important to note that director classification as "Successful" or "Unsuccessful" is based on their average ROI rather than total ROI. This classification method might yield instances where directors classified as successful could exhibit lower average profits than those deemed unsuccessful due to the equal weighting of movie ROI values.

Budget vs. Profit: The chart shows the spread of directors concerning their average budgets and profits, providing an insight into the range and distribution of financial performance across different directors.

Number of movies as Bubble Size: Larger bubbles would signify a director with more extensive experience or a higher volume of work.

This visualisation can assist in strategic decisions regarding director selection for upcoming projects and analysing the relationship between experience and financial success.



Actor metrics unveiled: Budget, Profit, and ROI analysis

This visual analysis encompasses approximately 400 actors interpreting multiple movies released in Great Britain from 2003 to 2022. To ensure clarity, actors engaged in at least in 5 movies were included, preventing chart overcrowding while offering a representative sample.

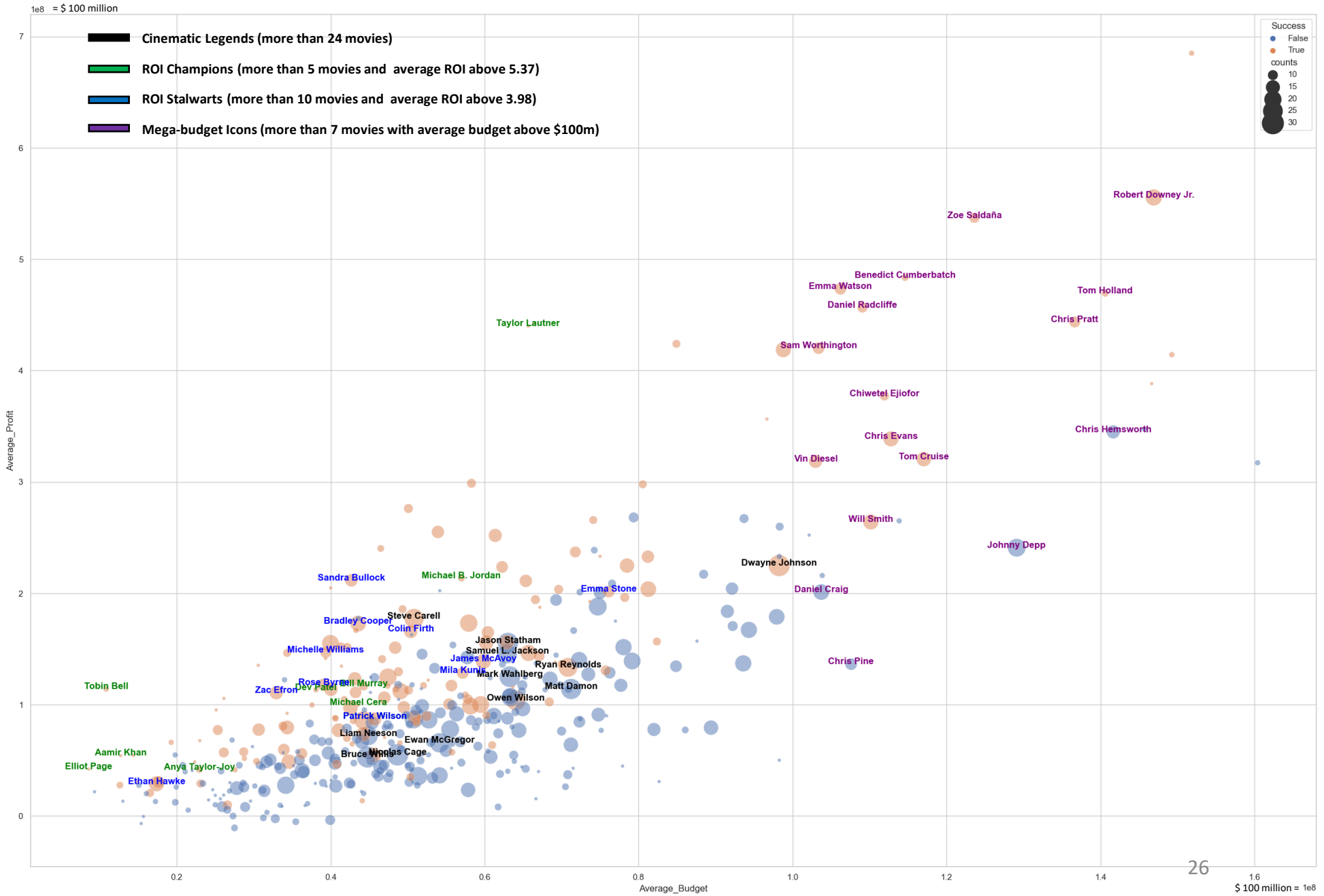
Each bubble on the chart represents an individual actor with its positioning indicating the average budget and average profit across their movie portfolio spanning the last two decades. The size of each bubble corresponds to the number of performances by the actor, while the colour categorizes actors based on whether their average ROI exceeds 2 or not. Differently coloured text highlights actors with diverse track records, ensuring clear differentiation and visibility.

It's important to note that actor classification as "Successful" or "Unsuccessful" is based on their average ROI rather than total ROI. This classification method might yield instances where actors classified as successful could exhibit lower average profits than those deemed unsuccessful due to the equal weighting of movie ROI values.

Budget vs. Profit: The chart shows the spread of actors concerning their average budgets and profits, providing an insight into the range and distribution of financial performance across different actors.

Number of movies as Bubble Size: Larger bubbles would signify an actor with more extensive experience or a higher volume of work.

This visualisation can assist in strategic decisions regarding actor selection for upcoming projects and analysing the relationship between experience and financial success.



Predictive Modelling

Introduction: Machine Learning overview

Machine Learning Overview

Machine learning involves the development of algorithms and models that enable computers to learn from data and make predictions or decisions without explicit programming. It encompasses various techniques to identify patterns, learn from data, and make predictions or decisions.

Classification in Machine Learning

Classification is a type of supervised learning where the goal is to categorise data into predefined classes or categories based on input features. It's used when the output variable is a category, such as "successful" or "unsuccessful" in this project. Classification algorithms learn patterns from labelled data to predict the class of new, unseen data.

Predictive Models in Identifying Successful Movies

Using machine learning models will allow us to analyse the movie-related features discussed in the Feature Engineering section of this report and understand how they correlate with a movie's success. These models will learn from historical movie data to predict whether a new movie is likely to be successful or not based on these features.

Benefits of Predictive Models

Insights Extraction: Predictive models help extract valuable insights by identifying which features or combinations of features are most influential in determining a movie's success. This information aids in understanding what factors contribute to a successful film.

Decision Support: They assist in decision-making by providing quantifiable predictions. For instance, these models can guide stakeholders in investment decisions, marketing strategies, or movie selection for production houses or studios.

Optimizing Resources: By predicting success, these models help optimise resource allocation. For instance, they can help in budget planning or targeting specific audience segments.

In summary, machine learning, particularly classification models, is instrumental in this project as it leverages historical movie data to predict the success of new movies based on various features. These models facilitate data-driven decision-making and provide valuable insights for stakeholders in the film industry.

Introduction: Algorithms and Ensembles

We implemented some of the most important binary classification algorithms and ensemble methods:

Logistic Regression

Linear model used to predict the probability that a movie belongs to a particular class (e.g., successful or unsuccessful) based on given features. We tried this algorithm because of its simplicity and effectiveness in cases where the **relationship between features and the probability of success is assumed to be linear or when the classes are separable by a linear boundary**.

K-Nearest Neighbours (KNN)

A non-parametric algorithm that classifies data points based on their similarity to neighbouring points. Think of movies as neighbours on a street; KNN calculates the proximity of a movie's features to its neighbouring movies and assigns the same class as its neighbours. KNN is considered due to its ability to **capture nonlinear relationships**, especially when **movies with similar characteristics tend to have similar levels of success**.

Support Vector Machines (SVM)

A powerful algorithm that finds an optimal hyperplane to separate data into different classes. SVM is utilised for **finding the best decision boundary** between successful and unsuccessful movies, **especially when there's a clear separation between the two classes**.

Decision Tree

They **segment the dataset into subsets** by posing a series of yes/no questions to classify movies, **based on the most significant features**. The algorithm is **highly interpretable** and adept at handling both numerical and categorical data.

Random Forest

An ensemble method that constructs multiple decision trees and combines their predictions. It effectively **mitigates overfitting**, helping to handle noise, **and manages a large number of features**, capturing complex relationships between them to predict movie success.

Extreme Gradient Boosting (XGBoost)

An ensemble method that sequentially builds trees, with each correcting the errors of its predecessor. It excels at handling complex interactions among features and works well with heterogeneous data. Gradient Boosting significantly improves accuracy by combining weak models in a sequential manner, **learning from previous mistakes, and building a stronger model iteratively**.

Voting Classifiers (Hard and Soft)

These classifiers combine predictions from multiple models. In Hard Voting, the final prediction is based on the majority vote of all individual classifiers. Soft Voting uses predicted class probabilities and averages them to decide the final class. **These classifiers can improve accuracy by considering diverse predictions from different algorithms**, such as the above.

Model development process

In the model development phase, we designed a robust process to build and evaluate predictive models for movie success classification. Here's a breakdown of our approach:

Data Split: Our dataset was divided into training (70%) and testing (30%) sets to facilitate model training and evaluation. The 70%-30% split aims to strike a balance between having sufficient data for training to build a robust model and having enough testing data to adequately evaluate its performance and evaluate overfitting.

Hyperparameter Tuning: : Leveraging 5-fold cross-validation, we fine-tuned hyperparameters for each algorithm within the training data. This approach bolstered the models' robustness and applicability to unseen data.

Handling Imbalance: Given the dataset's imbalance (with only 35% movies marked as successful), we experimented with various class weight options. Cross-validation guided us in identifying the most suitable weight distributions.

Metrics Consideration: We employed diverse scoring metrics—accuracy, precision, F1, ROC AUC, and average precision—to comprehensively assess model performance.

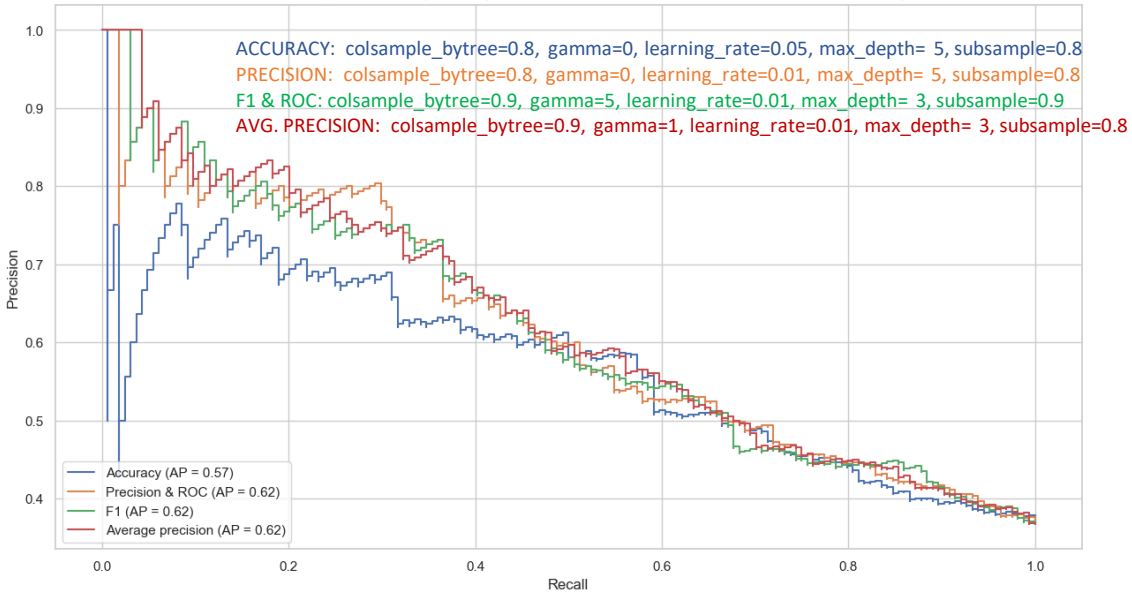
Precision-Recall Curve (PR Curve) Analysis: We plotted Precision-Recall curves for each optimised model within the testing data. These curves effectively visualise the trade-off between precision and recall for different models. Precision measures how many movies predicted to be successful are genuinely successful whereas Recall measures how many successful movies were accurately identified among all the successful movies.

Selection of Optimal Models: In our case, where only about 35% of movies are categorised as successful, using standard accuracy (total proportion of correctly classified instances) as evaluation metric might be misleading. Suppose our model predicts every movie as unsuccessful. In that case, it might still achieve high accuracy (around 65% since most movies are indeed unsuccessful), but it completely misses the point of identifying successful movies, which is crucial. The PR curve helps for the following reasons:

- It focuses on Positives: It specifically looks at how well the model performs on successful movies, which is more critical than how it performs on unsuccessful movies.
- Highlights Trade-offs: This curve showcases the trade-offs between precision and recall for different thresholds set by the model to predict success.
- Choosing the Best Model: A model with a higher area under the precision-recall curve is generally preferred. It shows that the model performs better at balancing precision and recall across different thresholds, even when dealing with imbalanced data.

The chart illustrates the models we evaluated for the XGBoost algorithm. Each model has different hyperparameters to optimise their respective performance metric. The PR curves help us to identify that the accuracy-optimised model is underperforming across various decision thresholds, warranting consideration of the other three models.

Example: Comparing XGBoost PR curves within the testing data



Customising decision thresholds

Users of our model will have the ability to tailor precision-recall ratios according to their specific preferences and goals. Here's how our model grants this flexibility:

Decision threshold customisation

- Our model offers the flexibility to adjust decision thresholds – the classification probability cutoffs.
- Lowering the threshold would increase recall but reduce precision, and vice versa.
- By altering these thresholds, users can decide whether they want to maximise precision, recall, or striking a balance between both.
- The ability to set decision thresholds puts control in the users' hands, ensuring tailored outcomes.
- This dynamic adjustment of decision thresholds ensures that our model remains adaptable, catering to diverse user requirements and facilitating better decision-making.

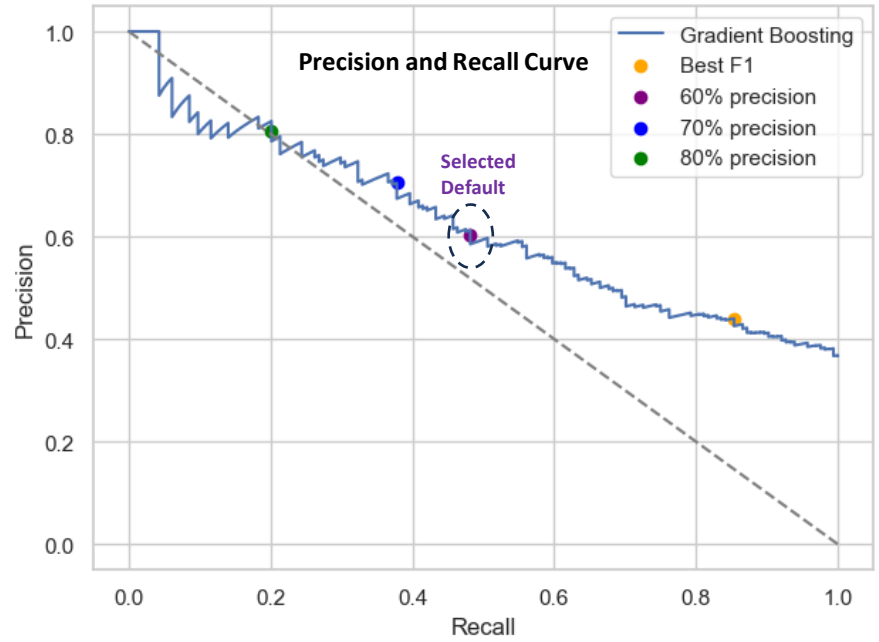
Visual reference for threshold adjustment

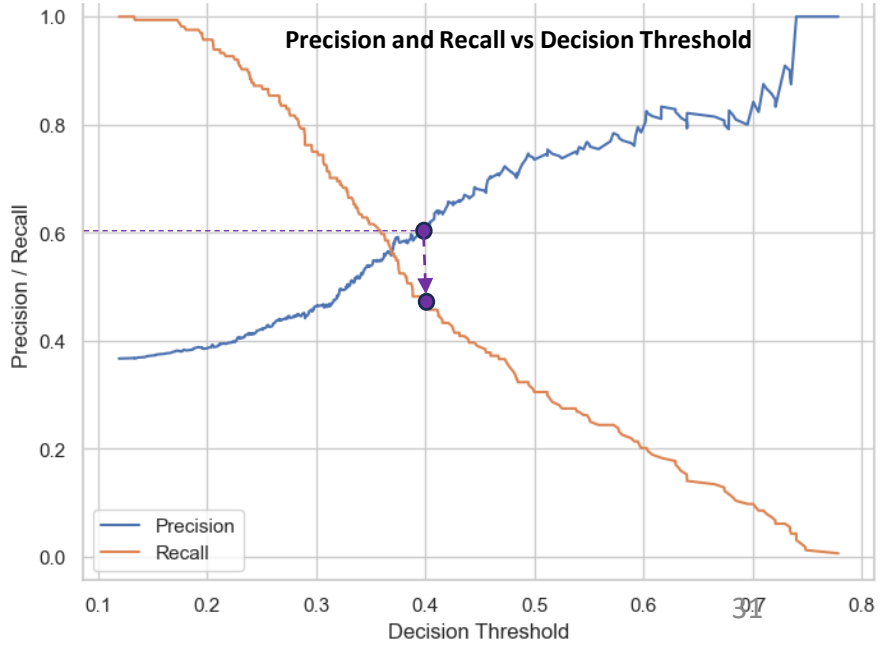
We've provided an example Precision-Recall curve chart with markers, showcasing various threshold selections and their respective precision-recall trade-offs.

Additionally, a chart illustrating Precision and Recall vs Threshold has been included for illustrative purposes to show the relationship between precision, recall, and threshold adjustments.

Default threshold setting

We've set a default decision threshold ensuring a minimum of 60% precision, aiming to assure studios' financial profitability (the current average precision rate for the 1,408 studios in our dataset is 29%), while offering ample investment opportunities. The model evaluation will be based on this default decision threshold.

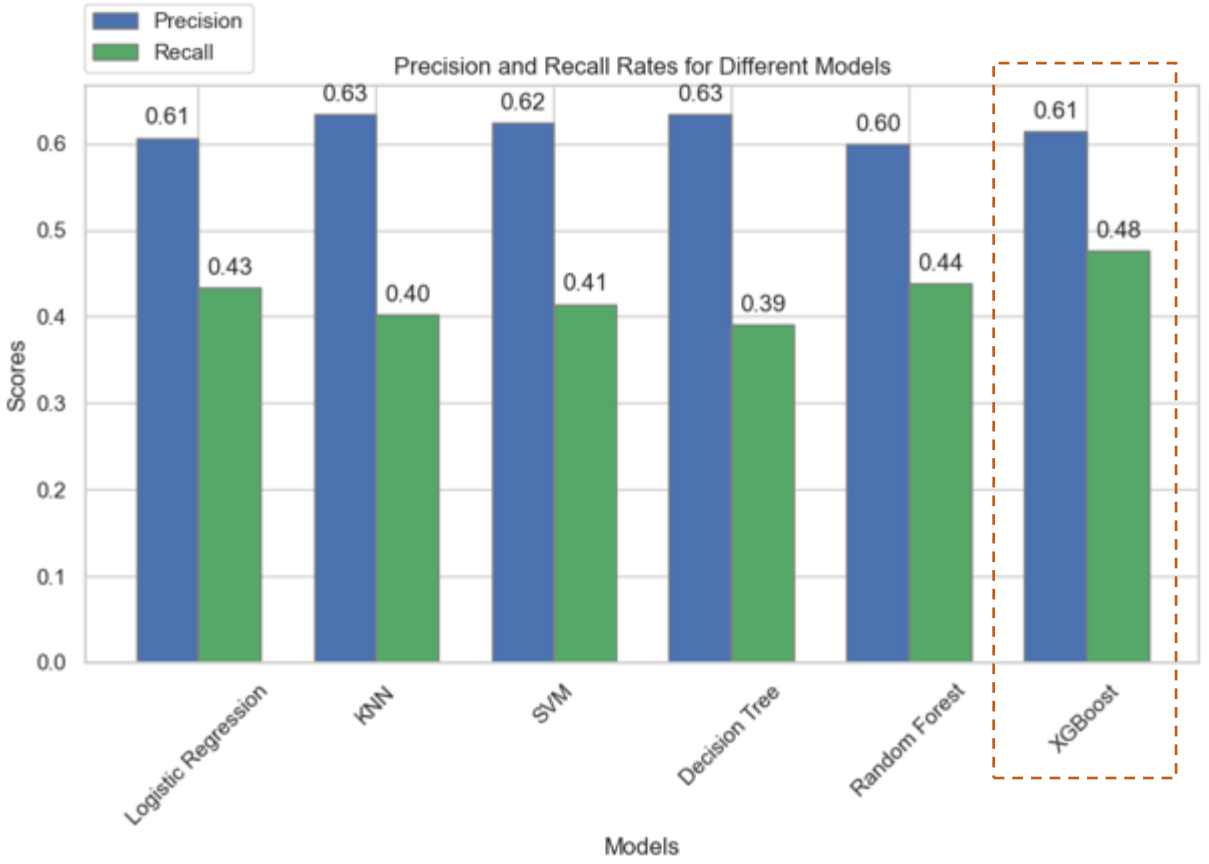


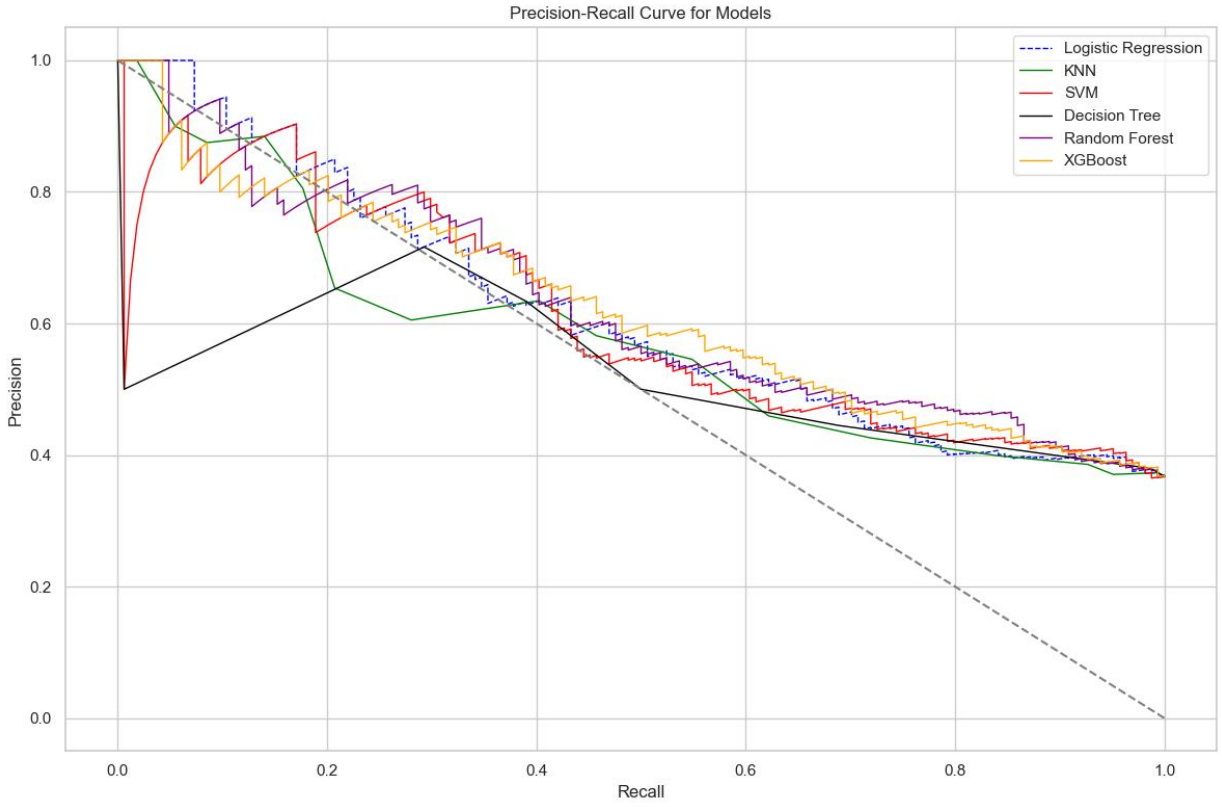


Predictive performance: All models

The line chart shows the precision-recall curves for each one of the algorithms implemented in this project. Most models deliver a balanced trade-off between precision and recall across various threshold values except for the KNN and Decision Tree models.

The grouped bar chart visually represents the precision and recall rates for the models utilised in our analysis. This comparison aids in understanding how different models perform in terms of correctly identifying successful movies (precision) and capturing the total number of successful movies (recall). At the selected threshold, the XGBoost algorithm is the best performer, delivering a precision in line with the other models but with a moderately higher recall rate.





Voting Classifiers

Voting classifiers combine predictions from multiple individual machine learning models, known as base models or estimators, to make a final prediction. They aggregate the predictions from diverse models and determine the final prediction based on majority voting (hard voting) or weighted average probabilities (soft voting) from these base models.

Here are some reasons why voting classifiers can be advantageous and outperform individual models:

Ensemble Learning: Voting classifiers belong to the ensemble learning category, leveraging the strengths of multiple models to compensate for their individual weaknesses. They combine diverse models to create a more robust and accurate final prediction.

Reduced Variance: By integrating predictions from multiple models, voting classifiers can reduce the variance of the final prediction compared to individual models. This reduction in variance often leads to a more stable and reliable prediction performance.

Improved Generalization: Voting classifiers can generalise better on unseen data by considering predictions from different models that have learned different patterns or features in the data. This diversity helps capture a broader spectrum of information from the dataset.

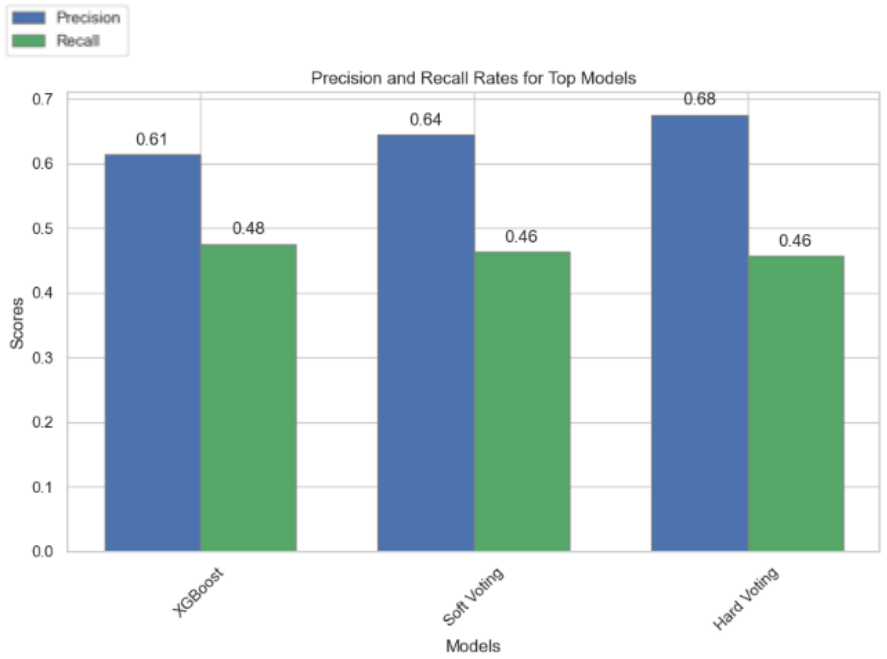
Mitigating Model Biases: If different base models have different biases due to their design or training data, voting can help offset these biases and lead to a more balanced and unbiased final prediction.

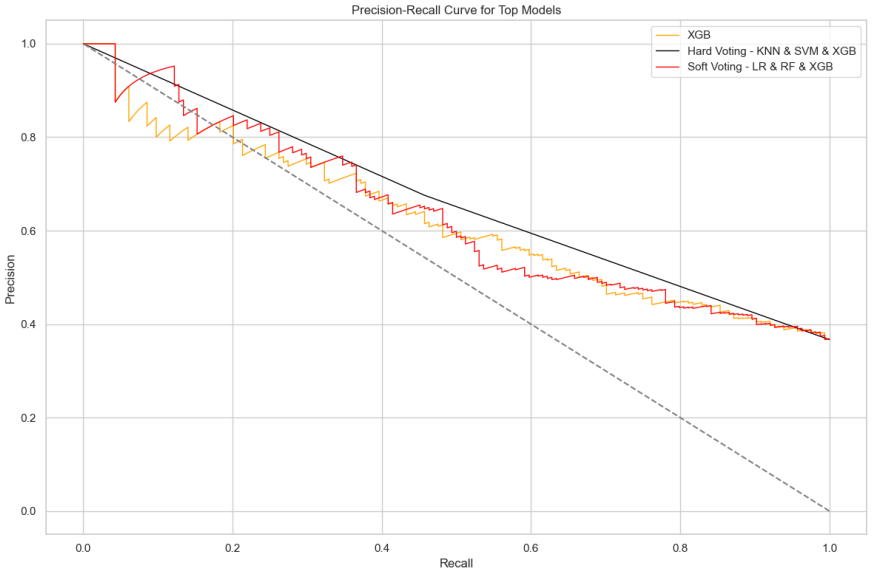
However, it's important to note that the effectiveness of voting classifiers relies on the diversity and quality of the base models. If the base models are highly correlated or if they all suffer from similar weaknesses, the voting classifier might not yield significant improvements.

We have used all the possible combination of models (61 combinations in total) for both the Soft and Hard classifiers and the two models that deliver the highest trade-off between precision and recall are the following:

Soft Voting: Logistic Regression + Random Forest + XGBoost

Hard Voting: KNN + SVM+ XGBoost





Multi-criteria decision making

Choosing the appropriate model for our movie success prediction depends on various factors beyond just the predictive performance metrics. We have considered the following aspects to make a more effective decision. We compare the three models that had the best predictive performance.

Criteria	Description	XGBoost	Soft Voting	Hard Voting	Reasoning
Performance	Precision and Recall rates.	6	6	7	The Hard Voting classifiers delivers a slightly higher precision with a similar recall.
Interpretability	How easily can stakeholders or end-users understand and trust the model's predictions?	6	5	4	XGBoost has moderate interpretability whereas the voting classifiers might be harder to understand considering they combine multiple algorithms.
Scalability / Efficiency	Computational resources required for training and inference.	7	6	5	XGBoost and Logistic Regression are generally efficient and scalable. KNN, Random Forest and SVM's efficiency decreases with large datasets.
Feature Importance (Explicability)	Understand which features contribute most to the predictions.	8	7	4	Logistic Regression, Random Forest and XGBoost provide feature importance scores whereas KNN and polynomial SVM don't.
Ease of Implementation and Maintenance	Ease of implementing the model into a production environment and maintenance considerations.	7	5	4	As a standalone model, XGBoost tends to be relatively easier to implement and maintain compared to voting classifiers.
	AVERAGE RATING	**6.8**	**5.8**	**4.8**	

The XGBoost algorithm, while demonstrating slightly lower overall performance compared to Soft Voting and Hard Voting classifiers, offers several significant advantages. It provides a clearer interpretation for stakeholders and higher explicability through accessible feature importance metrics. Additionally, XGBoost tends to be more scalable with larger datasets. Moreover, in terms of ease of implementation and maintenance in a production environment, XGBoost as a standalone algorithm is relatively simpler to manage compared to the complexity introduced by voting classifiers involving multiple models.

Considering these factors holistically, while XGBoost might exhibit marginally lower overall predictive performance compared to ensemble-based voting classifiers, its interpretability, scalability, explicability and ease of implementation and maintenance make it a robust and practical choice for predicting the success of new movie projects.

Results and Findings

Movie Success Predictor: Decision process

XGBoost combines many individual decision trees, each specializing in different patterns within the data. Each tree learns from the mistakes of the previous ones, improving the overall team's performance. This teamwork allows XGBoost to make accurate predictions by considering various aspects of the data and leveraging the strengths of different decision-making approaches.

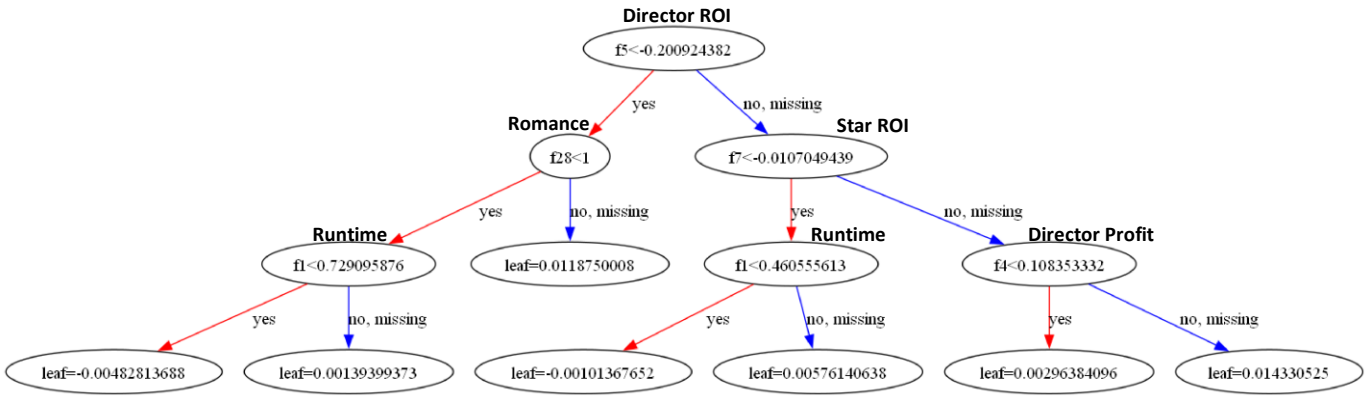
The MSP model consists of 300 decision trees. The visual diagram below illustrates one of these decision trees. Each node represents a split based on a particular feature and its associated threshold. The terminal nodes (leaf nodes), contain raw scores that can be converted to probabilities using a logistic function to interpret the likelihood of a movie belonging to a certain class.

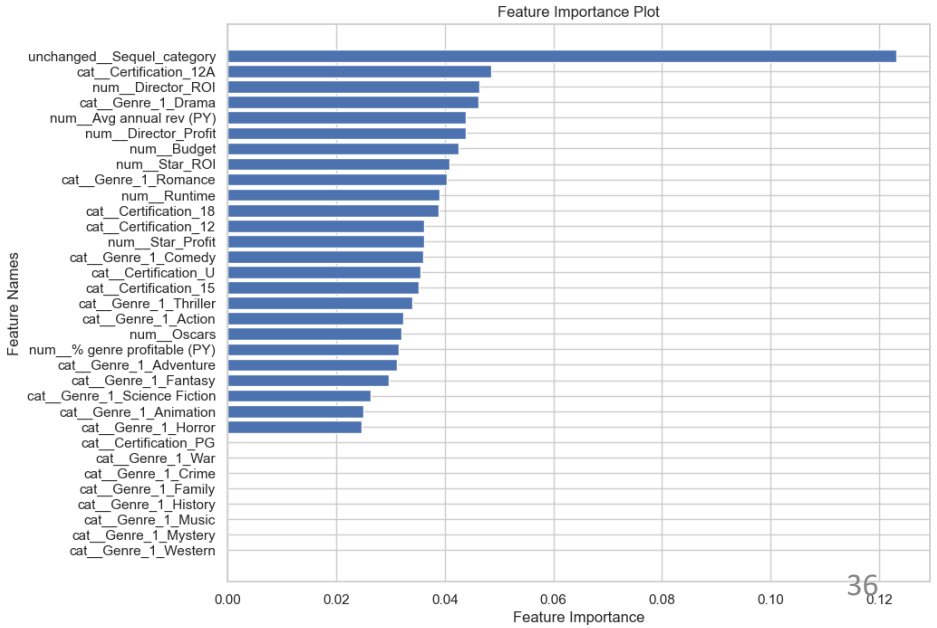
For prediction, the model checks at each node a specific feature and decides whether to follow the left or right branch based on a threshold value. For instance, in the illustration below, the model would check if the average Director's ROI of the movie is above or below 2 (that's the -0.2 threshold converted from its standardized value back to the original scale). Then the movie would be passed through the following nodes until reaching a leaf, which would assign a probability of success. In our example, a movie with an average Director's ROI below 2 and non Romantic genre would be assigned a 50% success probability (calculated from the sigmoid function applied to the 0.011875 raw score).

This process continues through the trees in the ensemble with each tree making a prediction based on its set of rules and features. The predictions from the trees would be combined, averaging the probabilities and using a threshold to determine the final predicted class for a movie.

The feature importance plot illustrates the contribution of each feature in making predictions. This plot ranks the features based on their importance scores, which represent their relative influence on the model's predictions. The importance scores are calculated using the average gain brought by a feature when used in decision trees across the ensemble. Success largely depends on being part of a movie Sequel, but the MSP model overly relies on this. Future approaches should further develop techniques like Feature Engineering, Feature Selection, and Dimensionality Reduction to reduce dependence on this single factor.

Illustration of an individual decision tree part of the MSP model





Movie Success Predictor : Outcomes

Confusion matrix

The confusion matrix depicts the performance of our MSP model by tabulating the predicted and actual classifications for a dataset. It helps in assessing the model's accuracy by breaking down the results into four categories based on the predicted and true classes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

We can conclude the following from the confusion matrix results and subsequent evaluation metrics:

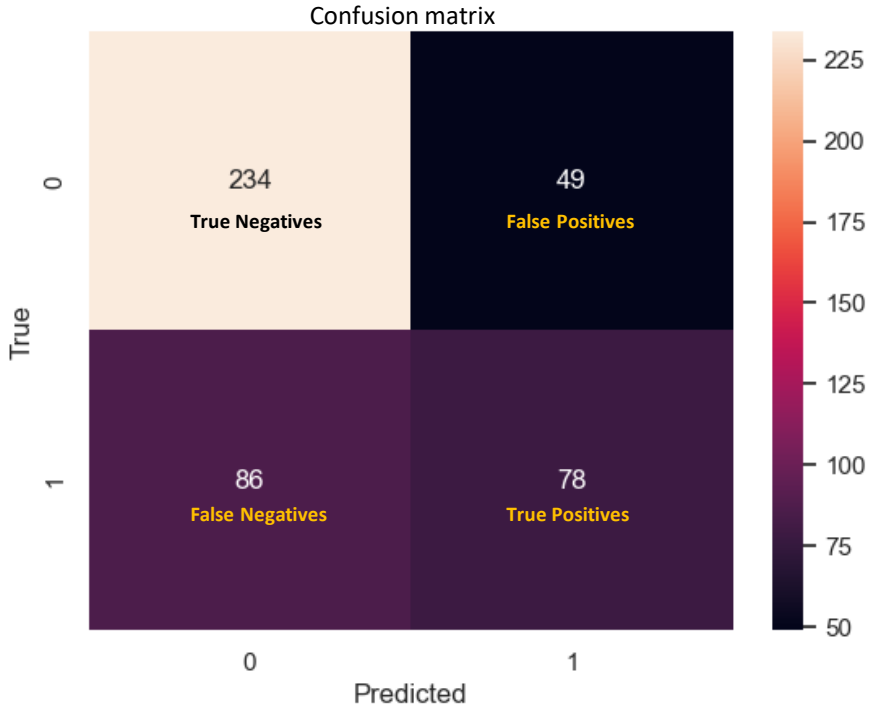
- The accuracy of the model stands at approximately 70%, showcasing its overall correctness in predictions.
- The precision of the model is moderately good, indicating that when the model predicts a movie as successful, it is correct around 61% of the time.
- The recall is relatively lower, suggesting that the model only captures approximately 48% of the actual successful movies.
- The F1 score, which considers the balance between precision and recall, is 54% and therefore shows that the model has room for improvement in finding a better balance between the precision and recall metrics.

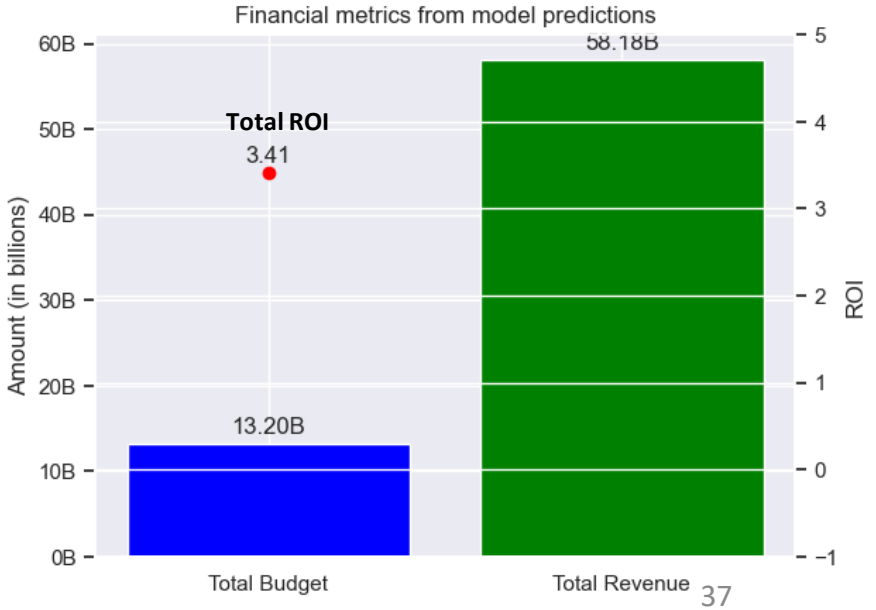
Financial performance

What would happened if we invested in the 127 movies recommended by our model? That's what the bar chart in this slide shows.

- The blue and green bars respectively show the sum of production budgets and revenues from movies recommended by the MSP. The sum of all the budgets is 13.2 billion, whereas the total revenue stands at 58.2 billion.
- The red point signifies the calculated total ROI resulting from these investments. In this scenario, if investments were made in all movies recommended by the MSP model, it would lead to a total ROI of 3.4. This ROI metric indicates a 340% increase from the initial investment in terms of profits, highlighting the potential financial gains achieved by following the model's recommendations.

How does this outcome compare with the industry? Let's explore it in the next slide.





Comparison with Industry Peers

This visualization compares the performance of our Movie Success Predictor (MSP) model against 110 studios that have released at least 5 movies in Great Britain between 2003 and 2022. The minimum has been selected to ensure we are comparing our model to relevant studios.

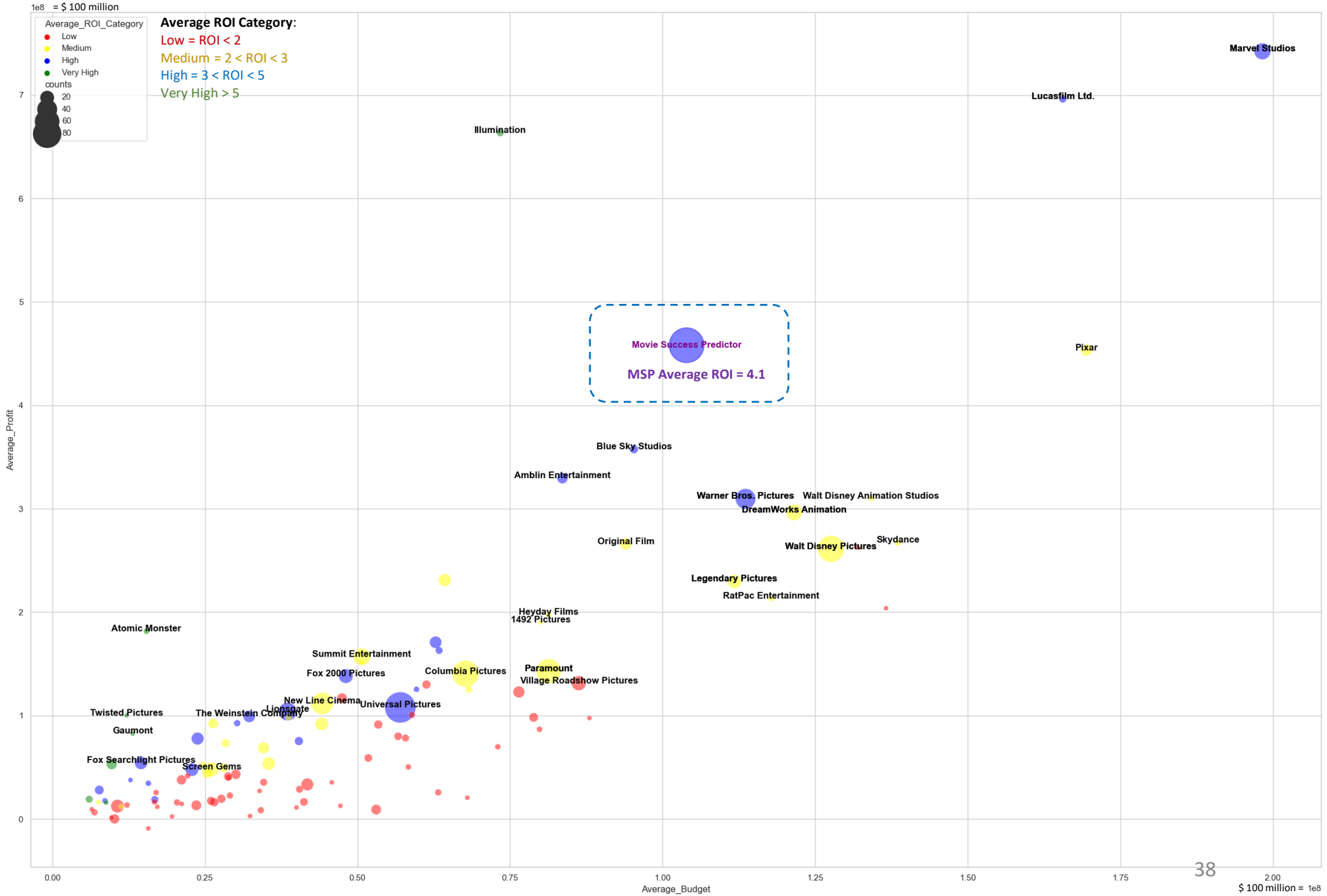
Each bubble in the chart represents a studio, with its position along the horizontal (x) and vertical (y) axes denoting the average budget and average profit, respectively. The size and colour of each bubble correspond to the number of movies produced and Average ROI category.

Budget vs. Profit: Employing average values helps to illustrate the effectiveness of studios in generating profit relative to the budget invested per movie.

Number of movies (counts) as Bubble Size: Larger bubbles represent larger studios with more activity.

Average ROI Category as Bubble Colour: It's important to note that Studios' ROI classification is based on Average ROI rather than Total ROI. This offers a better representation of the studio's typical movie performance, smoothing out the impact of outliers. This classification method might yield instances where studios classified in a higher category exhibit lower average profits than those in inferior categories due to the equal weighting of movie ROI values.

This visualisation aids to compare our model performance against industry peers. The MSP model, with an average ROI of 4.1, is at the 89th percentile.



Limitations and Conclusions

Limitations and Future work

Data quality

The dataset primarily sourced from The Movie Database (TMDB) API showcases generally good quality, but **deeper examination to validate accuracy and reliability** is advisable.

Measurement of financial metrics & Success definition

The calculated profit relies on estimated production budget and reported box office revenue, potentially obscuring the true profit due to varied accounting practices and revenue sources. While using ROI > 2 serves as a useful proxy for financial success, it might not capture all scenarios.

Defining success solely based on ROI > 2 might oversimplify success criteria. We could consider defining different success categories and using **Multiclass Classifiers** to predict these targets.

Imbalanced data

The dataset's 35% proportion of successful movies could affect the model's ability to accurately learn patterns related to success. This limitation has been addressed by adjusting the scale of the positive weights in the MSP model in order to place more emphasis on correctly classifying the minority class, potentially leading to an increase in Recall.

Feature selection

Our model heavily relies on using Sequel Category for predictions. Feature Engineering, Feature Selection and Dimensionality Reduction could be employed to **reduce the over-reliance on a single feature**.

Current features may not entirely capture influential factors for a movie's success. **Network-based features** such as average actor-director profitability were among the most significant features in a **research paper*** published in 2016 and hence these features could be considered to enhance the predictive power of our model.

Another interesting avenue to explore is whether Natural Language Processing (NLP) algorithms could be used to study the **predictive capabilities of movie scripts**. Additionally, external factors (e.g., **market trends, socio-economic events**) that could impact a movie's success should also be considered.

Correlation between past and future

Algorithms can be highly accurate as long as future patterns remain broadly similar to the past. Considering the evolving nature of the movie industry and audience preferences, it is crucial to **refine the model continually and engage with film industry professionals** to get their feedback on an ongoing basis.

Ethical practices

We could establish **ethical guidelines and governance principles** for developing our model, ensuring alignment with societal values and ethical standards. Additionally, more effort could be done to **strive for inclusivity** by considering a broad range of factors that could contribute to movie success beyond financial metrics.

* Lash, M. T., & Zhao, K. (2016). Early Predictions of Movie Success: The Who, What, and When of Profitability. Journal of Management Information Systems, 33(3), 874–903. https://doi.org/10.1080/07421222.2016.1243969

Conclusions

- Our Movie Success Predictor (MSP) model will use **Extreme Gradient Boosting (XGBoost)** to predict movie success (ROI equal or greater than 2).
- XGBoost's scalability, explicability and ease of implementation and maintenance make it a **robust and practical choice** for predicting the success of new movie projects. Its complex nature may limit interpretability but this can be mitigated by Feature Importance explanation, visualisations and overall effective communication.
- **Cross-validation and robust evaluation techniques** have been implemented to ensure the model's generalizability and reliability on unseen data.
- The model has been **adjusted to mitigate class imbalance** and give more weight to successful movies, improving minority class prediction.
- MSP allows for **flexibility in threshold settings based on stakeholders' preferences** or business needs to adapt to varying success criteria.
- While MSP showed promising results, it exhibited a **heavy reliance on the "Sequel" feature**. Future iterations will focus on Feature Engineering, Selection, and Dimensionality Reduction techniques to diversify feature importance and reduce dependency on this singular attribute.
- The model demonstrated reasonable performance metrics, with an Accuracy of 70%, Precision of 61%, Recall of 48%, ROC AUC of 65%, and PR AUC of 48%. Notably, the **model's precision rate of 61%** signifies a marked improvement compared to the industry's success rate. The industry's success rate stands at 29% across all studios and rises to 41% among the top 50 studios producing a larger number of movies.
- MSP recommended investing in 127 out of 447 movies in the testing set. Had these recommendations been followed, the projected Total Worldwide Box Office revenue would have been $58.2 billion against a $13.2 billion production budget, resulting in a **Total ROI of 3.4**, exceeding the success threshold.
- The **Average ROI, at 4.1 and within the 89th percentile**, showcases MSP's capability among studios producing more than five movies over the last two decades. Only around 10 out of the 110 studios we investigated delivered a better Average ROI than our model.
- The evolving nature of the movie industry requires ongoing refinement of predictive models. **Collaboration with industry professionals** for feedback and model adaptation remains crucial for sustained accuracy.
- The project acknowledges the need for **ethical guidelines and governance principles** to ensure alignment with societal values. **Striving for inclusivity and broader considerations in defining success** beyond financial metrics was highlighted for future enhancements.

References

References

1. GitHub Repository for Supporting Evidence and Code: https://github.com/asanzribas/The-Crystal-Ball-of-Cinema/blob/main/Predicting%20movie%20success.ipynb
2. Lash, M. T., & Zhao, K. (2016). Early Predictions of Movie Success: The Who, What, and When of Profitability. *Journal of Management Information Systems, 33*(3), 874–903. https://doi.org/10.1080/07421222.2016.1243969