

# **Final Report: Data Science Challenge**

Advanced Predictive Analysis and Visualization of  
Movie Data

Advait Gupta  
28 April 2024

## Executive Summary

This report documents the methodologies, visualizations, and outcomes of a project aimed at developing a predictive model to forecast the release year and potential genres of upcoming movies based on a director's past work. Utilizing a rich dataset of movie metadata, we applied rigorous data preprocessing, innovative feature engineering, and sophisticated machine learning techniques to build models capable of making these predictions with high accuracy. The report delves deep into the data handling processes, model training, evaluation metrics, and the insights gained through various visualizations, providing a holistic view of our analytical approach.

## Introduction

The objectives of this study are twofold:

- To visualize and analyze the relationships between various features within the movie metadata.
- To construct and validate models that can predict the release year and genres of a director's forthcoming movie.

## Data Description

The dataset comprises the following columns, which include both numerical and categorical data:

- Basic Information: Color, Director Name, Movie Title, Genres
- Facebook Popularity Scores: Director Facebook Likes, Actor 1-3

Facebook Likes

- Performance Metrics: Num Critic for Reviews, Num Voted Users, Num User for Reviews, Gross, Budget
- Temporal Data: Duration, Title Year

- Textual Data: Plot Keywords
- Miscellaneous: IMDB Score, Aspect Ratio, Movie Facebook Likes, Content Rating, Language, Country

## Data Preprocessing

### ● Missing Value Treatment

- Numeric columns like gross and budget were filled using the mean of their respective columns to maintain data integrity.
- Categorical and textual data, specifically plot keywords, were replaced with empty strings where missing to enable text analysis.

### ● Feature Extraction from Text

- Sentiment analysis was conducted on 'plot\_keywords' to derive a 'sentiment\_score' feature, which represents the overall emotional tone of the movie's plot, ranging from negative to positive. This was achieved using NLTK's Sentiment Intensity Analyzer.

## Feature Engineering

### ● New Features Introduced

- **Profitability:** Calculated as the difference between 'gross' and 'budget'.
- **Return on Investment (ROI):**  $(\text{Profitability} / \text{Budget}) * 100$ , representing the financial return per dollar spent.
- **Popularity Score:** An aggregate of Facebook likes from the director and the top three actors, intended to reflect the movie's social media presence.
- **Release Interval:** This feature represents the time gap between consecutive films by the same director. For directors with only one movie, this interval is set to -1, acknowledging their lack of historical interval data

which impacts the model's ability to predict future release timings accurately.

- **Handling Special Cases**

- For directors with only a single release, typical predictive modeling for intervals between releases becomes infeasible. To manage this within our dataset, such entries have 'next\_release\_year' and 'release\_interval' set to -1, indicating a unique situation.

## Predictive Modeling

- **Genre Prediction**

- A RandomForestClassifier was trained using features like ROI, profitability, and popularity score, among others, to predict multiple genres of a movie simultaneously. The model achieved an accuracy of **99.51%** on the test data, demonstrating its effectiveness in recognizing genre patterns.

- **Release Year Prediction**

- The prediction of the release interval between movies was handled by both RandomForestRegressor and GradientBoostingRegressor. An ensemble method averaging their predictions was employed to improve robustness and reduce potential overfitting. The Mean Squared Error (MSE) for these models on test data were **0.0633, 0.0449, and 0.0470 respectively**, showcasing high prediction accuracy.

# Visualization of Feature Relationships and Trends

## Detailed Visual Analyses

### 1. Popularity by Genre (Director Facebook Likes)

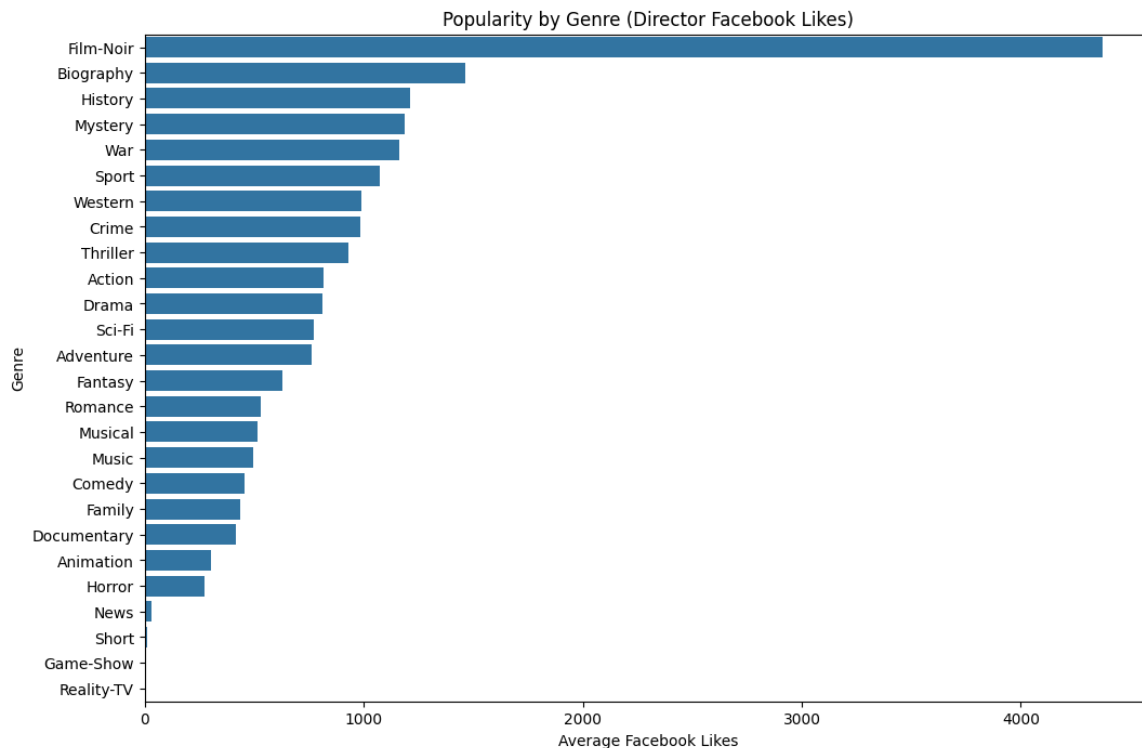
This visualization explores the relationship between movie genres and the average Facebook likes for directors associated with those genres. It provides an understanding of how the popularity of directors on social media varies across different genres. The metric used, average Facebook likes per genre, is calculated by aggregating the Facebook likes of directors for each genre and dividing by the number of movies in that genre. This measure gives insights into which genres tend to attract directors with higher social media engagement, potentially reflecting broader audience interest or marketing efforts.

#### Analysis:

- **Highly Popular Genres:** Genres such as Action and Drama may show higher average likes due to the prominence of directors in these genres who engage actively on social media.
- **Niche Genres:** Less mainstream genres might have fewer likes, possibly due to the niche audience or less emphasis on social media presence by the directors.

#### Visualization:

A bar chart represents each genre on the y-axis and the corresponding average Facebook likes on the x-axis. The genres are sorted by the average likes in descending order to easily identify which genres are associated with the most popular directors on Facebook.



### Insights:

- This analysis helps production companies understand which genres are currently popular and could be leveraged for marketing and promotional strategies.
- For potential directors or filmmakers, this visualization highlights which genres might offer more visibility based on current social media trends.

## 2. Average Gross by Genre

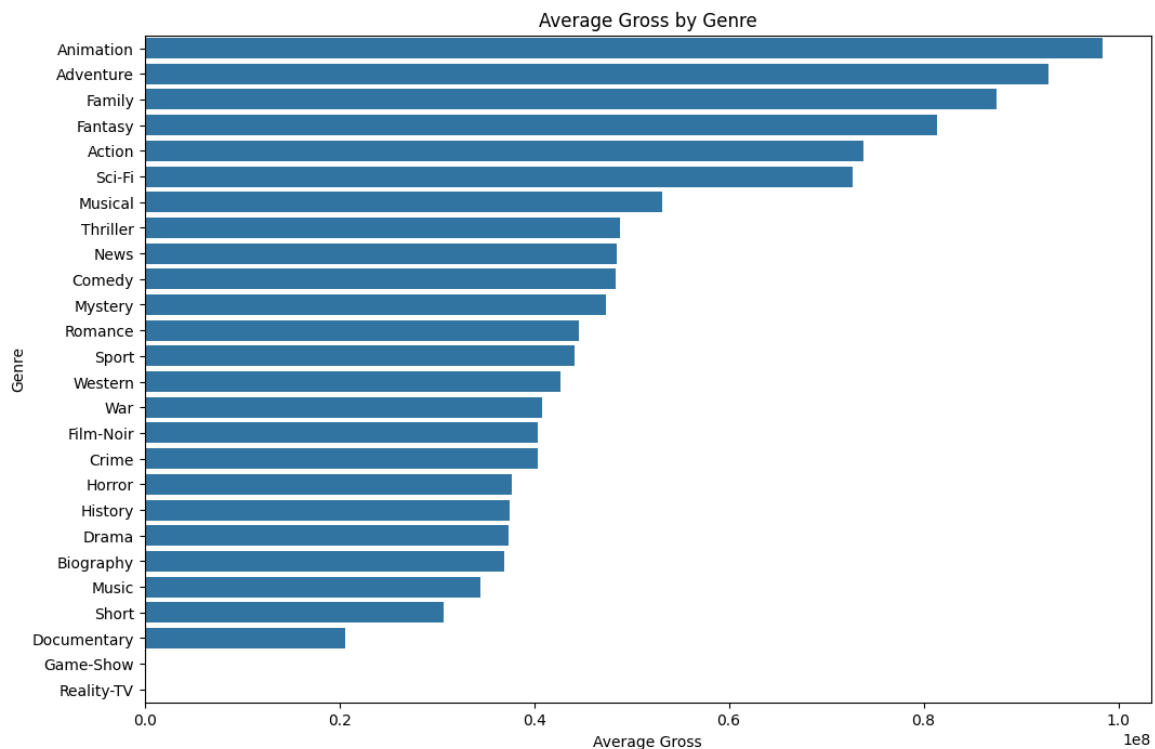
This plot displays the average gross revenue per genre, providing a straightforward metric of financial success across different types of movies likely to generate higher box office earnings. The average gross is calculated by taking the total gross for each genre divided by the number of films in that genre.

## Analysis:

- **Lucrative Genres:** Genres like Adventure and Action typically show higher average gross, reflecting their widespread appeal and large-scale production values.
- **Less Profitable Genres:** On the other hand, genres such as Documentary or Art House may exhibit lower average gross, indicating a more niche market and potentially lower investment and revenue expectations.

## Visualization:

A bar plot represents each genre along the y-axis with the average gross revenue depicted along the x-axis. The genres are ordered from highest to lowest average gross to clearly illustrate which genres are the most financially successful.



**Insights:**

- Production companies can use this data to align their production budgets and marketing efforts with the genres that promise higher returns.
- Investors and producers can assess risk and potential return based on genre performance, guiding strategic decisions in film production and financing.

**3. Director Influence on Gross**

This visualization examines how different directors influence the financial success of the films they helm. By plotting the average gross earnings against each director, particularly focusing on the top 20 directors by average gross, stakeholders can gauge which directors consistently deliver high box office returns.

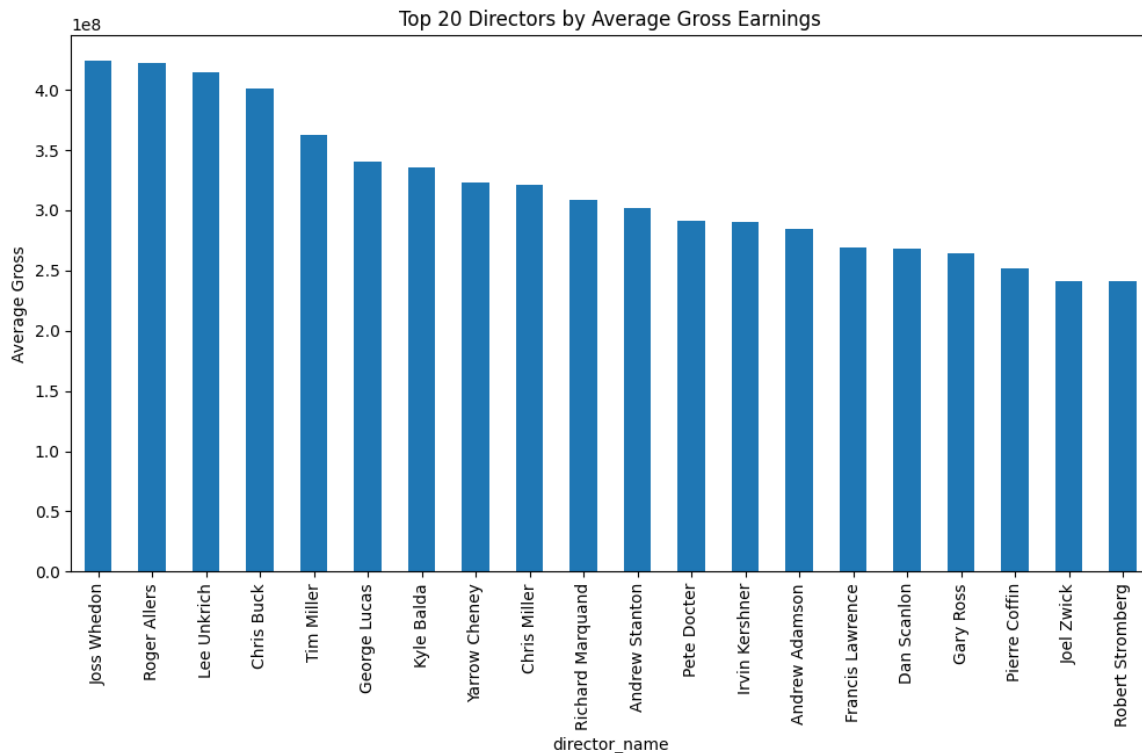
**Analysis:**

- **Top Performers:** Directors who consistently produce high-grossing films can be identified, highlighting their ability to attract audiences and generate revenue.
- **Market Trends:** Observing the directors leading in gross earnings can also reflect current market trends and audience preferences.

**Visualization:**

A bar chart displays the top 20 directors ranked by their average gross earnings. This visual helps to quickly identify which directors have the most significant financial impact in the industry.





### Insights:

- Studios might consider these directors for future projects, especially for genres that align with their proven track record.
- The visualization assists in recognizing directors who may have a 'brand' appeal that can be leveraged for bigger projects or promotional campaigns.

## 4. Temporal Trends in Movie Budgets and Gross

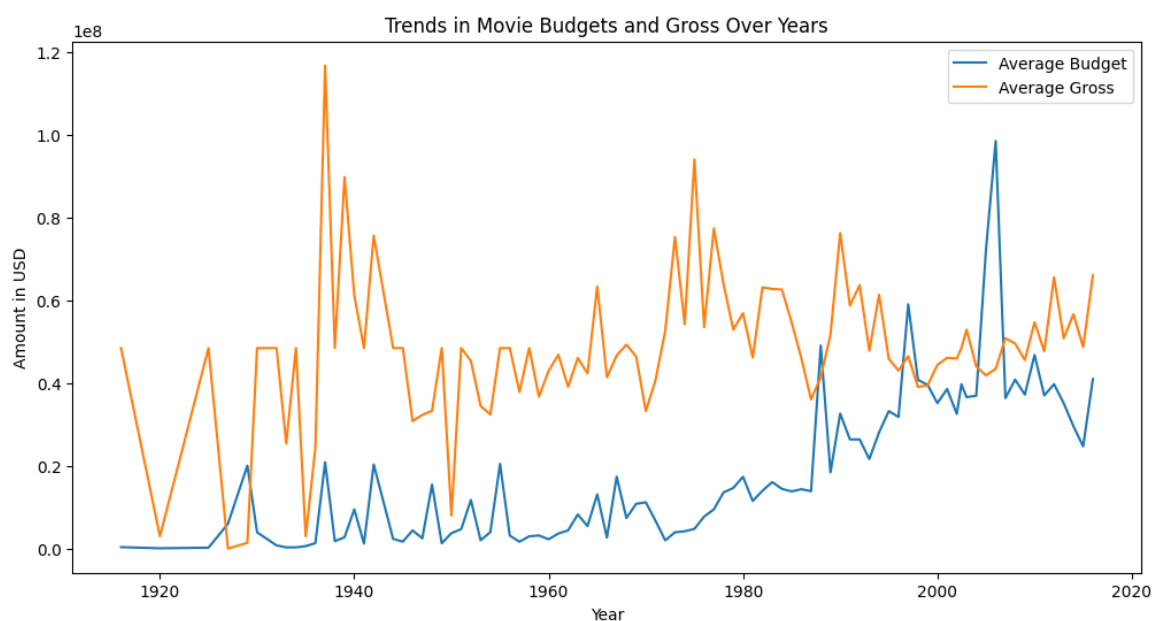
Tracking changes in average movie budgets and gross revenues over the years offers insights into the industry's financial health and shifts in movie production strategies.

### Analysis:

- **Budget Trends:** Increases in average budgets over the years might indicate rising production costs or more significant investments in high-quality content.
- **Revenue Trends:** Analysis of gross revenue trends can help understand whether the investments are paying off in terms of box office performance.

### Visualization:

A line graph shows the trends in average budgets and gross revenues over the years, with each year on the x-axis and dollar amounts on the y-axis. The lines for budgets and gross are labeled for clarity.



### Insights:

- This visualization can guide strategic planning for future investments in the film industry, providing a historical context for financial decision-making.
- Understanding these trends helps stakeholders anticipate future market conditions, potentially guiding more informed budgeting and financing decisions.

## 5. Cultural Impact: Average Gross by Language and Country

This section of the analysis delves into the cultural dimensions of film profitability, examining how the average gross revenue of movies varies by language and country. Such insights can reveal significant market differences and preferences that are influenced by cultural factors.

### Analysis:

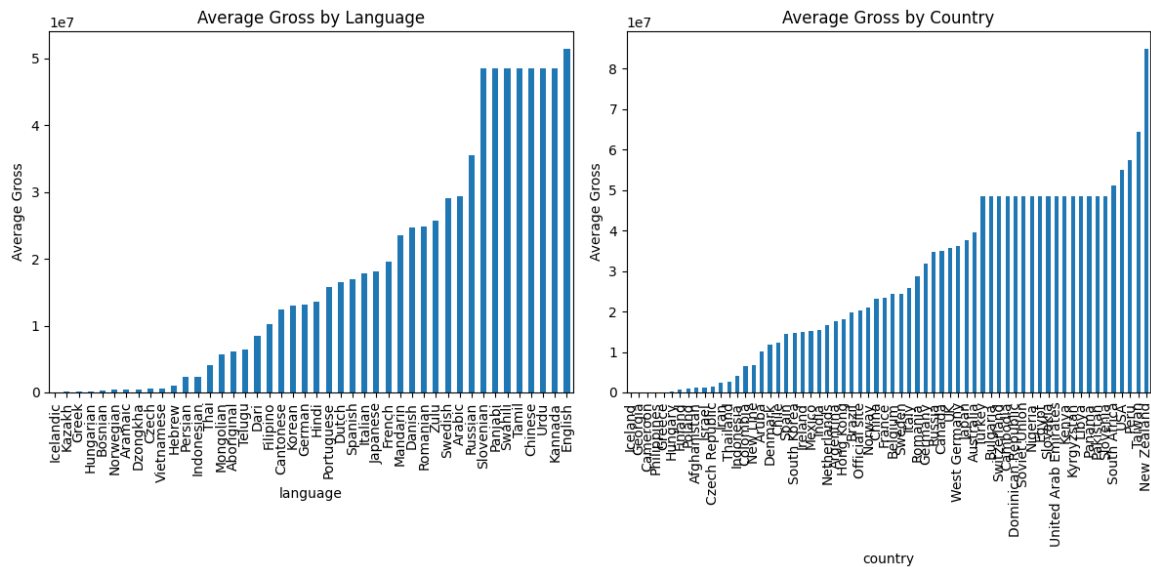
- **Language Influence:** Films in certain languages may generate more revenue, potentially due to a larger global audience or distribution in high-revenue regions.
- **Country of Origin:** The economic impact of movies can also vary significantly based on the country of production, which might reflect the country's film industry size, production quality, and international appeal.

### Visualization:

Two separate bar charts are created to illustrate these cultural impacts:

**Average Gross by Language:** A bar chart with languages on the x-axis and average gross on the y-axis. This visualization highlights languages associated with higher earnings, possibly indicating a wider international release or popularity.

**Average Gross by Country:** A bar chart with countries on the x-axis and average gross on the y-axis. It shows which countries produce films that perform well financially on a global scale, offering insights into successful film production hubs.



### Insights:

- Strategic Market Entry:** For producers and distributors, understanding which languages and countries yield higher returns can guide decisions about film production locations and language choices for dubbing or subtitles.
- Localization Strategies:** Knowing the financial performance of films by language and country can help studios tailor their marketing and distribution strategies to maximize revenue.

## Conclusion

This analysis highlighted several key insights:

- Directors' past box office performance and social media popularity are indicative of their future movie's financial success and genre.
- Certain genres consistently generate higher revenues, which can guide production decisions.
- The developed models are highly effective, as evidenced by the high accuracy and low MSE in predictions.

## Future Recommendations

- **Data Expansion:** Integrating more diverse datasets, including newer social media metrics and critic reviews, could further enhance the predictive power of the models.
- **Model Enhancement:** Experimenting with more advanced machine learning techniques and hyperparameter optimizations could yield even more accurate predictions.
- **Longitudinal Studies:** A continuous study involving data from new movie releases would help in refining the models over time.