# Movie Box Office Prediction

**Shuyue Li, Anish Rao, Liming Zheng**

## Abstract

The primary objective of this project is to develop a predictive model to estimate the final box office earnings of movies based on various features such as budget, running time, actors' and directors' past box office performance, awards nominations, release year, and IMDb scores. This system aims to provide insights to movie producers, investors, and marketers by forecasting the potential financial success of a film before its release.

## Database Design

Our database design aims to support the front-end client by enabling real-time data retrieval, addition, modification, and deletion. To meet these requirements, we chose MongoDB, a NoSQL database, due to its flexibility, scalability, and powerful query capabilities. This section outlines the key design decisions behind our database structure and the reasoning for each choice.

In order to perform preliminary normalization on the data, we split the data into separate collections, each serving a distinct purpose: storing basic movie information, movie performance data, user information, and feature usage data. This approach reduces data redundancy and enhances data consistency and integrity. In the movie_performance collection, for example, basic movie information is not duplicated but instead linked via the movie_id field, which references the _id from the movies collection. Similarly, in the feature_usage collection, user information is not duplicated, but the user_id field references the _id from the users collection. Additionally, data cleaning procedures, such as converting box office earnings to integers, IMDb scores to floats, and handling missing values (e.g., converting absent Oscar and Golden Globe award counts to 0), are implemented. By leveraging unique identifiers from other collections, we ensure clear relationships and traceability across the database, making it more efficient and ready for real-time operations. These normalization steps contribute to a well-structured and flexible database design.

**Schema Overview**

Our MongoDB database comprises four collections: movies, movie_performance, users, and feature_usage. Each collection stores specific types of data related to movies and user interactions. Here is a detailed description of each collection:
Movies Collection:

- _id: MongoDB automatically generates a unique ObjectId for each document.
- title: String – The movie title.
- director: String – The director of the movie.
- actors: Array of Strings – A list of actors in the movie (up to three in this case).
- genre: String – The genre of the movie.
- budget: Integer – The movie's budget (in dollars).
- release_year: Integer – The year the movie was released.
- imdb_score: Float – The IMDb score of the movie.
- description: String – A brief description of the movie (can be updated later).

Movie Performance Collection:

- _id: MongoDB automatically generates a unique ObjectId for each document.
- movie_id: ObjectId – The unique identifier of the movie, which links to the movies collection.
- final_box_office: Integer – The actual box office earnings for the movie (in dollars).
- earnings: Integer – The earnings from the movie.
- oscars_and_golden_globes_nominations: Integer – The number of Oscar and Golden Globe nominations the movie received.
- oscars_and_golden_globes_awards: Integer – The number of Oscar and Golden Globe awards the movie won.
- performance: String – A placeholder that tracks the difference between predicted and actual performance, such as box office prediction vs. actual earnings.

Users Collection:

- _id: MongoDB automatically generates a unique ObjectId for each document.
- username: String – The username of the user.
- email: String – The user's email address.
- full_name: String – The full name of the user.

- registration_date: String (Date format: "YYYY-MM-DD") – The date the user registered.
- last_login: String (Date format: "YYYY-MM-DD") – The last login date.
- preferred_language: String – The user's preferred language.
- role: String – The role of the user, either 'admin' or 'normal'.

Feature Usage Collection:

- _id: MongoDB automatically generates a unique ObjectId for each document.
- user_id: ObjectId – The unique identifier of the user, which links to the users collection.
- feature_name: String – The feature name (e.g., "box_office_prediction", "movie_trends", etc.).
- interaction_date: String (Date format: "YYYY-MM-DD") – The date of the interaction.
- interaction_details: String – A description of the user's interaction with the feature.
- prediction_outcome: String – The result of the prediction, which could be "success", "failure", or "N/A".

## Application Description

The application provides the following key features: regular users can search and view movie information and can also make box office predictions based on user-inputted features; admin users can perform create, read, update, and delete (CRUD) operations on the data. The frontend command line interface(CLI) offers various options for users to search movie information based on different input fields, such as movie title, runtime, etc. Once the user submits their input, the system automatically runs the corresponding query to retrieve data from the backend database and returns the results to the frontend interface, ensuring real-time and accurate information.

Additionally, admin users can input new movie information via the front end to add movie data. If any fields are missing, the system automatically fills in the missing information to maintain the completeness of the data. The system interacts with the backend MongoDB database, performing query operations to retrieve and update data, ensuring that all user actions are reflected in the database promptly, and maintaining data consistency.

## Data Collection

The data was acquired from a public dataset on Kaggle, named Movies Dataset[1], IMDB Dataset of Top 1000 Movies and TV Shows[2].

This dataset contains information about movies, such as movie titles, directors, actors, budgets, box office earnings, IMDb scores, etc. The data comes from multiple publicly available sources and has been organized and cleaned to ensure its accuracy and consistency.

During the data processing phase, the following steps were taken:

1. Data Cleaning: After reading the raw CSV files, we checked the completeness of the fields and handled missing values. For example, some movies had missing award data, so these empty fields were filled with zeros to ensure consistency across the dataset.
2. Data Transformation: Before importing the data into the database, certain fields were transformed. For instance, the budget and box office earnings fields were converted from strings to integers, and the IMDb score field was converted into a floating-point number.
3. Data Merging: We acquired movie description data from another CSV file (imdb_top_1000.csv) and updated the main movie dataset with the descriptions, using the movie titles as the matching condition

## Data analysis/Exploratory Data Analysis (EDA)

In this section, we present insights derived from an exploratory analysis of the dataset, leveraging visualizations to uncover relationships and trends in the data. These analyses aim to provide a deeper understanding of the factors influencing box office performance.

### Relationship Between Budget and Box Office Earnings

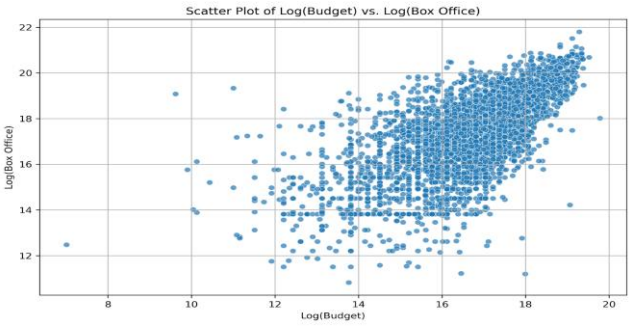


Figure1

The scatter plot of log-transformed budget versus log-transformed box office earnings demonstrates a strong positive correlation between these variables. As the budget increases,

box office earnings tend to rise proportionally on a logarithmic scale. However, there are some outliers where movies with modest budgets achieved significantly higher earnings, potentially due to other factors like strong storytelling, critical acclaim, or marketing success.
This observation underscores the importance of budget allocation in determining a movie's commercial success, though it is not the sole determinant.

### Distribution of IMDb Scores

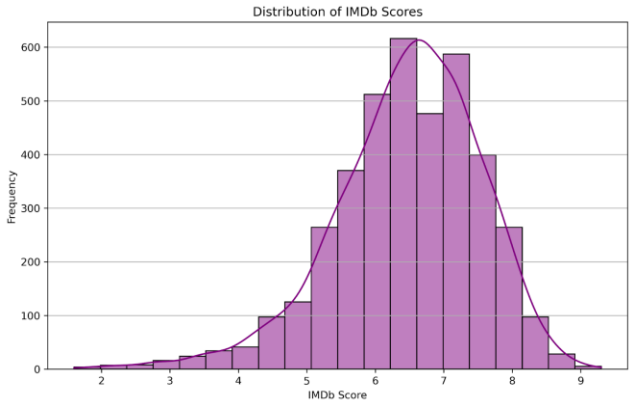


Figure 2

The histogram of IMDb scores reveals a slightly left-skewed distribution, with most movies scoring between 6 and 8. This indicates that the majority of movies in the dataset received moderate to high ratings from viewers. Very few movies fall below a score of 4 or above a score of 9, reflecting the general quality distribution of the sample.
This insight is essential for understanding the general audience reception and its potential influence on box office performance.

### Top Performers in the Dataset

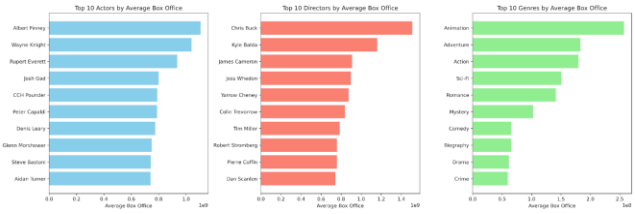


Figure 3

The bar charts highlight the top actors, directors, and genres by average box office earnings:

- **Top Actors:** Among the top-performing actors, Albert Finney and Wayne Knight lead in average box office earnings, showcasing their participation in commercially successful films. This suggests that certain actors bring significant value to a movie's financial success, either through their performance or marketability.
- **Top Directors:** Directors like Chris Buck, James Cameron, and Kyle Balda rank highly in terms of box office performance, indicating that directors with a track record of successful films often deliver strong commercial results.
- **Top Genres:** The animation genre is the most successful, followed by adventure and action. These genres often have broader appeal, catering to both younger and older audiences, which contributes to their higher earnings.

## Machine Learning Model

In this project, we used a Random Forest Regressor to predict movie box office earnings. The reason for choosing this algorithm is its ability to handle a large number of features and its good fitting capacity, which can capture non-linear relationships in the data. Furthermore, Random Forest is known for its high accuracy and robustness, making it well-suited for regression tasks involving complex features and large datasets.

We first preprocessed the data, which included filling missing values, handling outliers, standardizing the data, and performing feature selection by random forest model feature importance tool.

Then we did feature engineering and created feature importance chart.

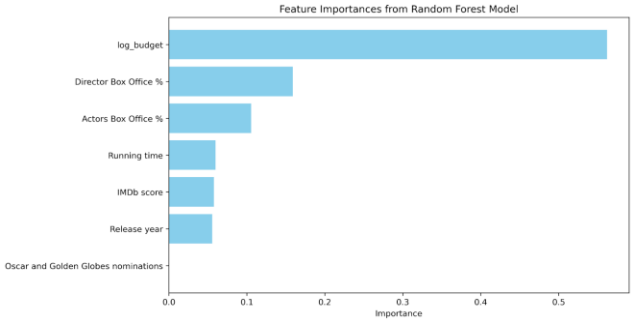


Figure 4

Based on the feature importance analysis, the most significant predictor of box office earnings is the log-transformed budget (log_budget), which accounts for approximately 56.2% of the model's explanatory power. This indicates that the budget plays a crucial role in determining a movie's box office success. The second most important feature is the director's box office percentage (Director Box Office %), contributing around 15.9%, followed by the actors' box office

percentage (Actors Box Office %), which accounts for approximately 10.6%. These two features suggest that the involvement of high-profile directors and actors has a noticeable impact on a movie's performance at the box office.

Other features such as running time (Running time), IMDb score, and release year (Release year) contribute relatively smaller amounts, with running time having a slightly higher importance than IMDb score and release year. Interestingly, the Oscar and Golden Globes nominations feature has zero importance, implying that in this particular model, nominations do not significantly affect box office earnings. This insight highlights the model's focus on financial and talent-related aspects rather than awards recognition when predicting box office performance.

Next, the data was then split into training and testing sets, with the training set used to train the model and the testing set used to evaluate its performance.

After training the model, we obtained feature importance rankings, which helped identify the features most influential for predicting box office earnings. We also used GridSearchCV for hyperparameter tuning to select the best model configuration.

The model evaluation was based on two key performance metrics: Mean Squared Error (MSE) and $R^2$ Score.

| | Before | After | Improve by % |
|---|---|---|---|
| Mean Squared Error | 0.9274255 | 0.8966250 | 3.44% |
| R^2 Score | 0.6043658 | 0.6175051 | 2.13% |

Figure 5

Additionally, the model achieved an accuracy of 88.83% (the percentage of predictions with an error less than 10%), demonstrating that the model is highly accurate in predicting box office earnings.

Lastly, the model reveals that key factors such as budget, actors' box office percentage, director's box office percentage, and IMDb score significantly influence a movie's box office performance. Movies with higher budgets and well-known actors or directors generally perform better at the box office. Additionally, films with higher IMDb scores tend to attract more viewers. Oscar and Golden Globe nominations also have a positive impact, as they increase a movie's visibility. These insights highlight the importance of both financial and reputational factors in determining a movie's commercial success.

# Reports

### Report 1: What We Found

This project developed a predictive model to estimate movie box office earnings based on features like budget, actors' and directors' past performance, IMDb scores, and more. Data analysis revealed a strong positive correlation between budget and box office earnings, although some lower-budget films achieved high earnings. Additionally, the involvement of well-known actors and directors, along with the movie genre, significantly impacted box office performance, especially in animation. The machine learning model showed that budget, actors' and directors' box office percentages, and IMDb scores are key predictors of box office earnings, while Oscar and Golden Globe nominations had minimal impact. These findings provide valuable business insights for filmmakers and investors.

### Report 2: Trends in Box Office Performance Over Time

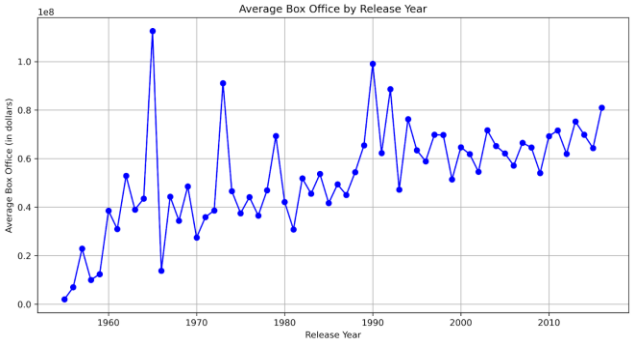


Figure 6

The line chart showing average box office earnings by release year illustrates a steady upward trend over the decades. This increase could be attributed to factors such as inflation, larger budgets, advancements in visual effects, and growing global accessibility to movies.
However, there are fluctuations that may correspond to specific economic conditions, technological breakthroughs, or the release of highly successful blockbuster franchises during certain periods.

**Report 3: Actual vs. Predicted Box Office**

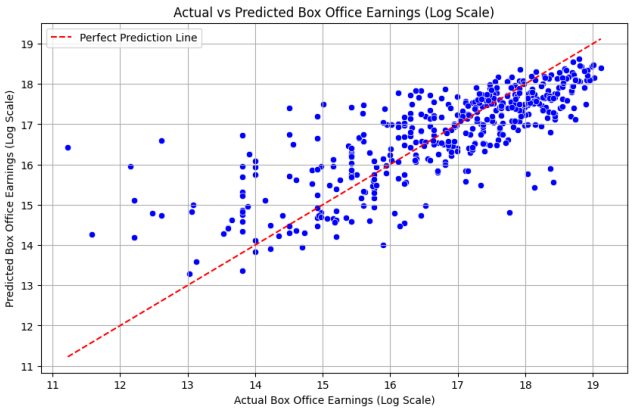


Figure 7

Based on the scatter plot of actual vs. predicted box office earnings (on a log scale), the majority of the data points align closely with the "perfect prediction line" (the red dashed line), indicating that the model's predictions are generally accurate. Most of the points are concentrated between the actual box office earnings of 17-19 (log scale) and predicted earnings of 16-18 (log scale). This suggests that the model performs well in predicting box office earnings for movies in this range. However, there are a few outliers that deviate significantly from the predicted line, indicating that for some movies, the model's predictions were not as accurate. Despite these outliers, the overall trend shows a strong positive correlation between the actual and predicted values, demonstrating the model's effectiveness in capturing the underlying patterns in the data.

**Report 4: Correlation Analysis Among Features**

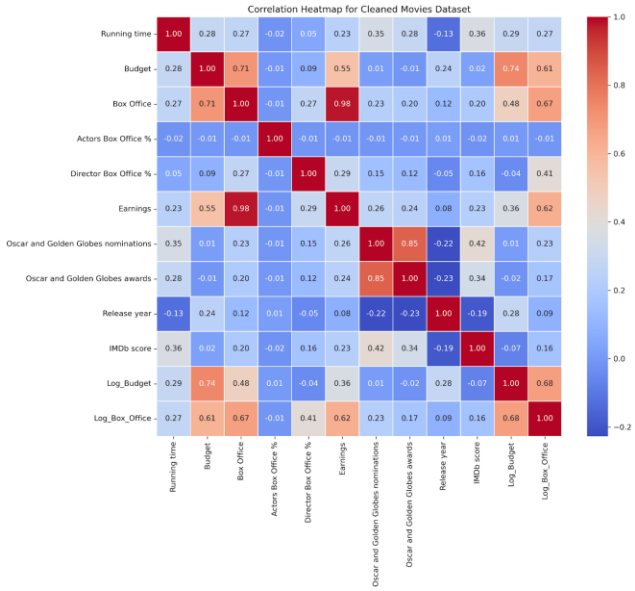


Figure 8

The correlation heatmap for the cleaned movies dataset provides a visual representation of relationships among different features. Key findings include:

1. High Correlation Between Budget and Box Office Revenue:

The budget shows a high correlation with box office revenue, with a coefficient of 0.71. This indicates that higher budgets are strongly associated with higher box office earnings, likely due to better production quality, increased marketing expenditure, and broader distribution.

2. Moderate Correlation Between IMDb Score and Box Office Revenue:

IMDb scores have a moderate correlation with box office revenue, with a coefficient of 0.42. This suggests that higher-rated movies tend to achieve better commercial success, though other factors also play significant roles.

3. Director's Box Office Percentage and Total Box Office Revenue:

The director's box office percentage correlates with overall box office revenue with a coefficient of 0.41, indicating that renowned directors contribute positively to a film's commercial performance.

4. Low Impact of Oscar and Golden Globe Nominations and Awards:

Oscar and Golden Globe nominations and awards have low correlations with box office revenue (0.23 and 0.17, respectively). While these accolades may enhance visibility and critical acclaim, they have a limited direct impact on earnings.

5. Higher Correlation in Log-Transformed Variables:

Log-transformed budget and box office revenue have a stronger correlation of 0.74, showing that log transformation reduces the influence of extreme values and highlights clearer relationships among variables.

**Report 5: Insights from Data Analysis**

From the exploratory data analysis, several key insights about the factors influencing box office performance emerge:

1. Key Drivers of Box Office Success:

Budget remains the most critical factor, with a strong positive correlation with box office earnings. Additionally, the involvement of prominent actors and renowned directors substantially boosts a movie's commercial performance. Certain genres, particularly animation, adventure, and action, also stand out as significant contributors to box office success, likely due to their universal appeal and ability to attract a wide range of audiences.

2. Trends Over Time:

The steady increase in average box office earnings over the decades reflects the growth of the movie industry, driven by advancements in technology, larger production budgets, and the globalization of film markets. Despite periodic fluctuations due to economic conditions or blockbuster releases, the overall trend highlights the increasing scale and impact of modern film productions.

3. Audience Reception and IMDb Scores:

Most movies in the dataset have IMDb scores clustered in the 6–8 range, suggesting that audiences generally favor well-made films. These higher-rated films tend to perform better financially, highlighting the importance of quality storytelling, production, and audience satisfaction in achieving box office success.

These insights emphasize the importance of financial resources, talent involvement, and audience perception in shaping a movie's commercial outcomes, providing valuable guidance for industry stakeholders.

## Conclusion

This project has provided valuable insights into the factors influencing movie box office performance and developed a robust predictive model using machine learning techniques. By analyzing various features such as budget, IMDb scores, and the involvement of renowned actors and directors, we identified key drivers that contribute to a movie's commercial success. Among these, budget emerged as the most significant predictor, followed by the influence of high-profile talent. Despite the importance of awards like Oscars and Golden Globes, our analysis showed that their impact on box office earnings was minimal, suggesting that factors like marketability and financial investment hold more weight.

The Random Forest Regressor model demonstrated high accuracy in predicting box office earnings, achieving an 88.83% accuracy rate, which confirms its potential as a tool for filmmakers, investors, and marketers. The model's feature importance analysis also revealed that, beyond financial factors, the roles of directors and actors have a substantial influence on a movie's success. This understanding is essential for stakeholders in the film industry, as it helps prioritize investments and optimize marketing strategies.

If given more time, we would have expanded the model to incorporate additional features such as marketing spend, distribution channels, and global market performance, which could further enhance the accuracy of predictions. Furthermore, exploring other machine learning algorithms, such as Gradient Boosting or XGBoost, could potentially improve the model's performance and robustness.

For future students of the DS 5110 course, we recommend starting early on data cleaning and feature engineering, as these steps are crucial for building a solid foundation for machine learning models. It's also important to continuously evaluate model performance through rigorous validation and hyperparameter tuning. Lastly, being open to exploring different algorithms and comparing their results to find the best solution for the problem is essential.

This project has been a rewarding learning experience, offering valuable insights into the dynamics of data analysis and machine learning, as well as their practical applications in the entertainment industry.

## References

[1] Oliva, D., "Movies Dataset," Kaggle, 2021. Available at: https://www.kaggle.com/datasets/delfinaoliva/movies.

[2] Shankhdhar, H., "IMDB Dataset of Top 1000 Movies and TV Shows," Kaggle, 2020. Available at: https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows.