Movie Prophet: Determining Movie Parameters to Predict Box Office Revenue

Name	Email	JHED
Mao Yang	myang112@jh.edu	myang112
Jiayu Pan	jpan49@jh.edu	jpan49
Rozanne Lim	rlim8@jh.edu	rlim8
Fan Yu	fyu24@jh.edu	fyu24

1. Introduction

For a long time, the cinematic landscape has served as a testament to human creativity and storytelling. The film production industry has become a cultural force that shapes our perceptions, influences our emotions, and captures the essence of various narratives. Beyond its artistic importance, the film industry is a massive economic engine, generating billions of dollars annually and providing employment opportunities for a vast network of professionals. However, a major ongoing challenge is the unpredictability of film success. Despite substantial investments in production budgets, star casts, and cutting-edge technology, not all films manage to secure a place in audiences' hearts or generate considerable profits. This enigma raises a crucial question: What are the underlying characteristics that transform films into commercial triumphs and cultural phenomena?

Existing work focuses on building a personalized machine model for the prediction of profits. Ahmed et al. conducted a comprehensive survey identifying three primary data mining approaches in movie revenue prediction: classification, regression, and clustering. Building on this foundation, researchers developed the Cinema Ensemble Model, which introduced a novel feature derived from transmedia storytelling theory and demonstrated superior performance compared to previous machine learning approaches. Our work builds on these foundations by combining multiple machine learning models, such as the Linear Regression Model, Generalized Additive Model (GAM), and Support Vector Machines with a Radial Kernel (SVM), to create a machine learning dashboard. Additionally, we created interactive data visualizations to provide information about how audiences’ tastes change over time and how different parameters relate to profits.

2. Objectives

The core problem we address in this project is the inherent uncertainty surrounding a film's box office performance. Utilizing a comprehensive dataset of American films from 2000-2025, which is enriched with diverse features such as production details, cast and crew information, genre

classifications, release timing, IMDb ratings, runtime, and critical scores, we aim to uncover the elusive traits that contribute to a movie's profitability and popularity. Our project focuses on identifying key characteristics that correlate with a film's success and determining the most effective machine learning models for predicting a movie's box office performance before its release. By doing so, we aim to provide filmmakers, studios, and industry stakeholders with valuable insights that can inform strategic decisions and mitigate the inherent risks associated with movie production.

3. Data Collection

3.1 Data sources

There were three main sources of data for this project. The first source of data was Wikipedia's List of American Films released in 2000 to 2025 (25 individual webpages), where the titles of movies was the data point of interest. The R package `rvest` was used to scrape the title of movies from tables on each Wikipedia page. Since the format of each page and URL was similar, this allowed the creation of a function and the use of map2_dfr function from the `purr` package to extract all the movie titles from each individual webpage from 2000-2025 at once, without running individual sections of code. The final list consisted of 6730 movies, after removing movies to be released after November 18, 2025 (data collection was conducted before that date).

The second source of data was the Open Movie Database (OMDB). This is a free, open-source RESTful service API that contains information on a variety of movies and associated information such as genre, director, and cast. An OMDB API wrapper for R, ROMDB, was used which allows for interaction with the API from R IDE to request the desired information. The find_by_title() function from the ROMDB package was used to obtain information on genre, box office revenue, IMDB rating, Metascore and runtime. The final dataframe consisted of 4757 movies, after removing movies that were not released in the desired time frame (2000-2025), and de-duplication of entries by IMDB ID.

The final source of data was the priceR package, which includes extraction of relevant inflation and exchange rate data from the World Bank API, and inflation adjustment calculations. Since the primary parameter of interest is box office revenue, inflation over time had to be accounted for in order to make direct comparisons across movies. The adjust_for_inflation() function from the priceR package was used to adjust the box office revenues to 2024 USD.

3.2 Data description

The final dataframe had 4757 movies and 30 columns. Relevant parameters included the title of the movies, the MPA film rating, date of release, genre, metascore, IMDB rating, and box office revenues (inflation-adjusted). Other variables were either variables that were cleaned (e.g., using

the `stringr` package to remove commas and convert to numeric type), or not used (e.g., IMDB ID, or awards). Original columns were kept as a point of reference. The figure below shows a subset of the final dataframe.

Figure 1: Subset of the final dataframe

Title str	Rated fct(7)	Released Date	Genre str	Genre1 fct(18)	Genre2 fct(22)	Genre3 fct(19)
Next Friday	R	2000-01-12	Comedy	Comedy	NA	NA
My Dog Skip	PG	2000-03-03	Comedy, Drama, Family	Comedy	Drama	Family
Play It to the Bone	R	2000-01-21	Comedy, Drama, Sport	Comedy	Drama	Sport
Supernova	PG-13	2000-01-14	Horror, Sci-Fi, Thriller	Horror	Sci-Fi	Thriller
The Boondock Saints	R	2000-01-21	Action, Crime, Thriller	Action	Crime	Thriller
Down to You	PG-13	2000-01-21	Comedy, Drama, Romance	Comedy	Drama	Romance
The Big Tease	R	2000-01-28	Comedy	Comedy	NA	NA
Isn't She Great	R	2000-01-28	Biography, Comedy, Drama	Biography	Comedy	Drama
Simpatico	R	2000-02-04	Comedy, Crime, Drama	Comedy	Crime	Drama
Gun Shy	R	2000-02-04	Comedy, Crime, Romance	Comedy	Crime	Romance
Scream 3	R	2000-02-04	Horror, Mystery	Horror	Mystery	NA
The Beach	R	2000-02-11	Adventure, Drama, Roman…	Adventure	Drama	Romance
Snow Day	PG	2000-02-11	Adventure, Comedy, Fami…	Adventure	Comedy	Family
The Tigger Movie	G	2000-02-11	Animation, Adventure, C…	Animation	Adventure	Comedy
Boiler Room	R	2000-02-18	Crime, Drama, Thriller	Crime	Drama	Thriller
Hanging Up	PG-13	2000-02-18	Comedy, Drama	Comedy	Drama	NA

3.3 Challenges

There were several challenges associated with data collection. The first challenge was in finding publicly available data on box office revenue. The IMDB website has an official API with all of the relevant parameters that we were interested in, but it was not free. Additionally, websites such as Numbers and Box Office Mojo had box office revenue data, but webscraping was not allowed on those websites. We were able to find open-source data from OMDB API, which we ended up using as a data source in the project.

The second challenge encountered was the limitations in the OMDB API and by extension, the ROMDB wrapper. The OMDB API did not include a function to obtain all movies released between 2000-2025; the closest function was to use the title of the movies or IMDB ID. Since IMDB ID was significantly more challenging to obtain, the decision was made to scrape movie titles of American films released from 2000-2025 from Wikipedia. The list of movie titles were then used to get information from the Open Movie Database.

Finally, there was a limitation with the number of API calls that could be made in one day. The free version of the OMDB API was limited to 1000 calls/day. As such, the initial list of 6730 movies was split up into smaller subsets, obtained the information from the API, and saved the data. The final dataframe was constructed by binding the separate, smaller, dataframes together.

Data cleaning and preprocessing posed a slight challenge. Many of the variables were not in the right format for analysis, or included special characters that needed to be removed. There was also a small amount of missing values for box office revenues, and those movies were removed from the dataframe since revenue was the primary variable of interest. De-duplication of movies according to their unique IMDB ID was also conducted, since the OMDB API produced duplicate records of the same movies.

4. Programming Paradigms

4.1 Functional programming paradigm

A functional programming style was used for the project by breaking down large functions into smaller, easily human interpretable functions. The functional programming paradigm was used primarily in data collection, data cleaning and processing, and data analysis. The first step of data collection was to create a list of movie titles from 2000-2025 from Wikipedia. However, the movies were listed on separate webpages by year. Instead of running 25 separate chunks of code for each different URL, a function was created to read in the URL links, extract movie title information using the CSS selector for Wikipedia tables, and return one dataframe. Additionally, a function was also created to query the OMDB API, since the built-in function only accepts one value at a time. The map_2dfr function from the 'purr' package was used to apply the OMDB API function to a list of movie titles, and return a dataframe. See **Figure 1** for an example of the final dataframe.

Functions such as mutate(), filter(), pull(), and sort() from the 'dplyr' package were used throughout the data cleaning and processing stage. The main processes for data cleaning were removal of special characters, changing the variable type, and inflation adjustment for the box office revenues. Removal of special characters was achieved using the 'stringr' package and other functions from the 'dplyr' package, while changing variable types used mutate() and functions from the 'forcats' package, for factor variables. For inflation adjustment, the 'priceR' package was used to get inflation factors for the years of interest (2000-2025). Since the goal was to get inflation factors for the 24 years, a function was created to get the inflation factors, and return a dataframe. The inflation factors were then used to adjust the box office revenues.

Figure 2: Function to retrieve inflation adjustment factors for years 2000-2024

```r
# Get unique years from the data
years_needed <- year_data %>%
  mutate(Year = as.numeric(Released_year)) %>%
  filter(!is.na(Year), Year >= 2000, Year <= 2024) %>%
  pull(Year) %>%
  unique() %>%
  sort()

# Get inflation factors for each year
# Create inflation lookup table
inflation_factors <- data.frame(
  Year = years_needed,
  Factor = map_dbl(years_needed, function(y) {
    adjust_for_inflation(1, from_date = y, to_date = 2024, country = "US")
  })
)

saveRDS(inflation_factors, "data/inflation_factors.RDS")
```

During data analysis and visualization, functions were developed to create interactive plots and tables. For instance, functions were created for the interactive movie table, genre heatmaps, and seasonal analysis of movie releases. By creating functions for these visualizations, it allowed for easier interpretability of each line of code, and the ability to easily make changes as needed, either to the data source or to the format and aesthetics of the plots. Furthermore, the goal was to create a dashboard with a consistent color scheme and formatting, and using functions to generate the plots made it easy to achieve that consistency.

4.2 Machine learning paradigm

4.2.1 Overview of the Machine Learning Pipeline

This project applies supervised machine learning techniques to predict movie box office revenue based on movie characteristics such as genre, release timing, popularity indicators, and production-related features. The modeling pipeline consists of five key stages:

(1) Data preparation and feature engineering

(2) Dataset construction and train–test splitting

(3) Model training using multiple algorithms

(4) Model evaluation using quantitative metrics

(5) Scenario-based prediction and interpretation

4.2.2 Data Preparation and Feature Engineering

4.2.2.1 Target Variable Transformation

The primary prediction target in this project is movie box office revenue (BoxOffice_num). Raw box office values are highly right-skewed, with a small number of blockbuster movies dominating the distribution. Directly modeling this variable would lead to unstable training and disproportionate influence of extreme values. To fix this issue, the target variable is

log-transformed: $\boldsymbol{y = \log(\text{BoxOffice\_num})}$

This transformation reduces skewness, stabilizes variance, and allows error-based metrics such as Mean Squared Error (MSE) to reflect relative prediction accuracy across movies with vastly different revenue scales.

4.2.2.2 Data Cleaning and Missing Value Handling

Observations with missing box office revenue or missing release year were removed, as these variables are essential for both modeling and feature construction. For remaining missing values, structured imputation strategies were applied: Aggregated statistics (e.g., average director revenue) were imputed using median values. Count-based features (e.g., number of prior movies by a director or actor) were imputed with a default value of 1. Temporal features (e.g., release month or quarter) were imputed using representative values to preserve seasonal structure. All numeric variables were checked for infinite values arising from logarithmic transformations or division operations. Infinite values were replaced with missing values and subsequently imputed using medians to ensure numerical stability across all models.

4.2.2.3 Feature Engineering

Feature engineering was performed to translate raw movie metadata into structured predictors that capture temporal patterns, popularity dynamics, and content characteristics. We apply extensive feature engineering to capture temporal, genre-based, popularity-related, runtime-based and creative factors that influence box office performance.

Temporal features were constructed to model long-term trends and seasonal release effects, including years since 2000, release month and quarter, as well as indicators for summer and holiday releases.

Genres were encoded using binary indicator variables to allow models to capture multi-genre effects, while genre count was included to represent genre complexity.

Audience engagement was captured using IMDb-related variables:

log_votes, votes_per_year, rating_squared,

These features help differentiate genuinely popular movies from those that have simply accumulated votes over time.

Both threshold-based and polynomial runtime features were constructed to capture non-linear effects of movie length, while sequel indicators were extracted from title patterns.

Historical performance statistics of directors and lead actors were aggregated and merged into the dataset, allowing the model to incorporate creative track records as predictive signals.

4.2.2.4 Dataset Construction and Train–Test Split

After feature engineering, a modeling dataset was constructed by selecting the engineered predictors and the transformed target variable. Rows with remaining missing values were removed to ensure consistent model inputs.

The dataset was split into training and testing subsets using an 80/20 ratio:

Training set (80%): used for model fitting

Test set (20%): reserved exclusively for performance evaluation

A fixed random seed was applied to ensure reproducibility. This separation prevents information leakage and ensures that evaluation metrics reflect true generalization performance.

4.2.3 Machine Learning Models

Linear Regression: Linear regression serves as a baseline model, assuming a linear relationship between predictors and log-transformed box office revenue. While simple and interpretable, this model is limited in its ability to capture non-linear effects and complex feature interactions.

Generalized Additive Model (GAM): The Generalized Additive Model extends linear regression by allowing smooth, non-linear functions of predictors:

$$\log(\text{BoxOffice}) = \beta_0 + \sum s_j(x_j)$$

Spline-based smoothing functions enable the model to capture non-linear relationships while maintaining interpretability. Smoothing parameters control model flexibility and reduce overfitting risk.

Random Forest: Random Forest is an ensemble learning method based on decision trees. Each tree is trained on a bootstrapped sample of the data, and predictions are averaged across trees. This model captures complex non-linear interactions and is robust to multicollinearity. However, it sacrifices interpretability and requires careful feature selection and evaluation.

Support Vector Machine (SVM): The Support Vector Machine with a radial basis function kernel projects input features into a higher-dimensional space to model non-linear relationships.

SVMs are powerful for complex patterns but sensitive to hyperparameters such as kernel width (gamma) and regularization strength (cost). Improper tuning may lead to overfitting.

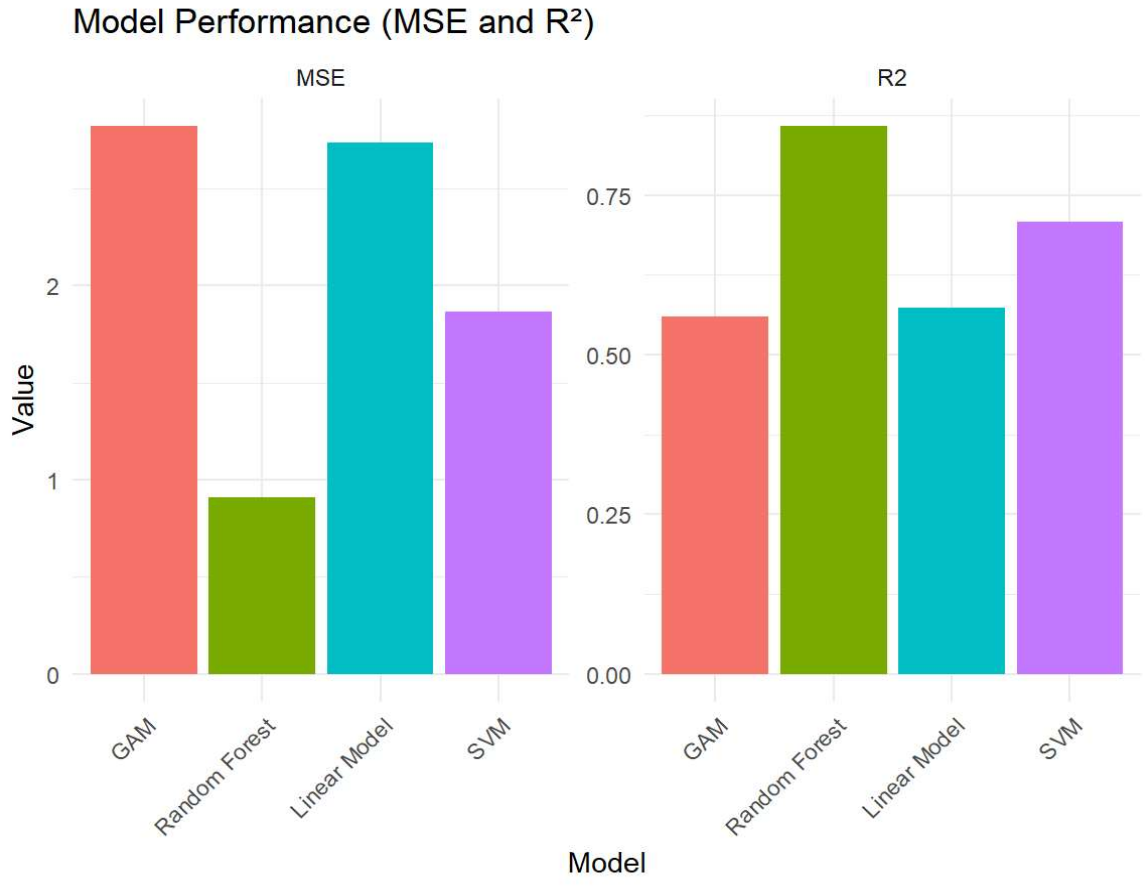
4.2.4 Model Evaluation

Two quantitative metrics were used for model evaluation:

Mean Squared Error (MSE): Measures average squared prediction error. Lower values indicate better predictive accuracy.

R^2 (Coefficient of Determination): Measures the proportion of variance in log box office revenue explained by the model. Higher values indicate better explanatory power.

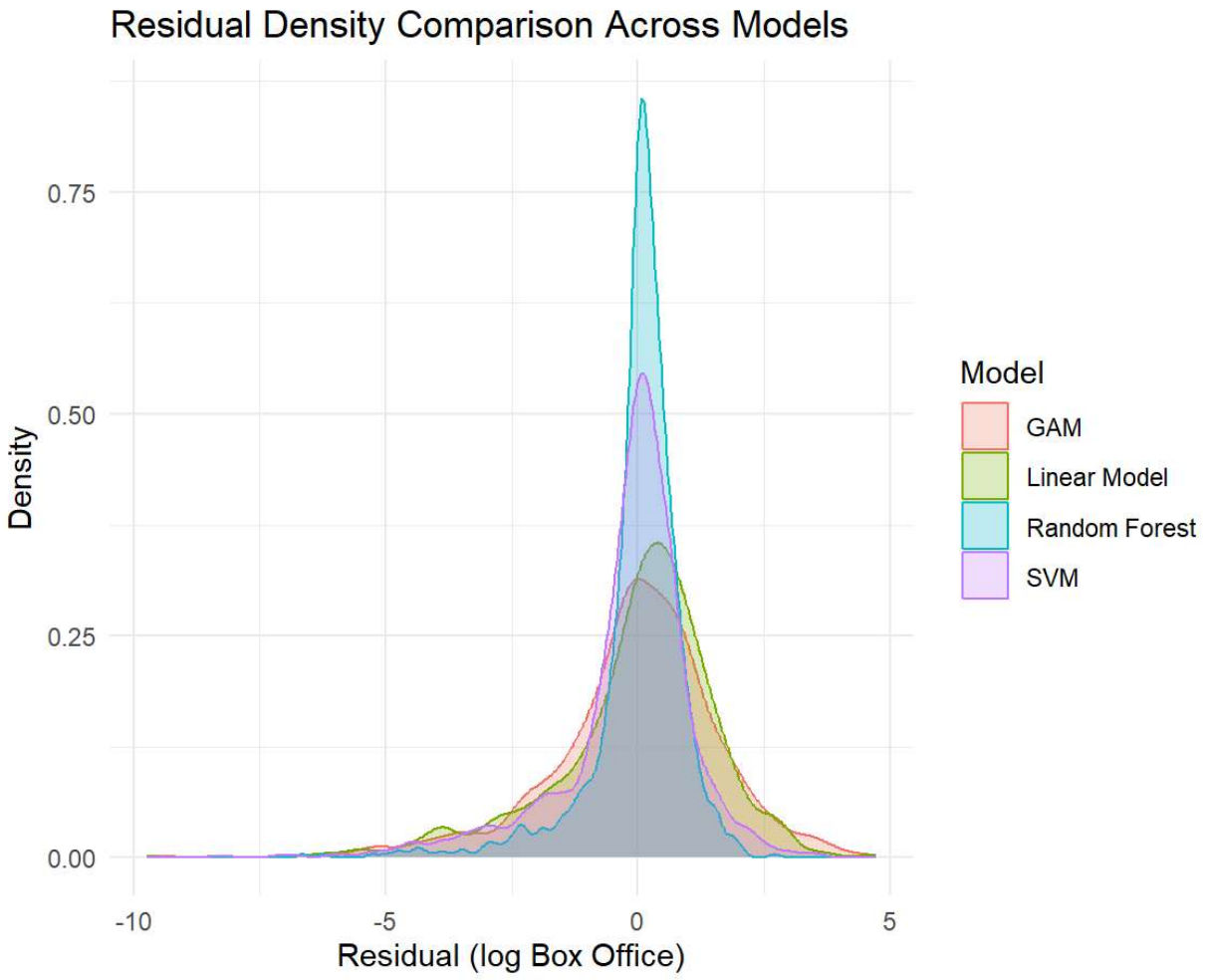
Figure 3: Comparison of Model Performance Using Mean Squared Error (MSE) and R^2



This figure compares the predictive performance of four machine learning models—Generalized Additive Model (GAM), Linear Regression, Random Forest, and Support Vector Machine (SVM)—using Mean Squared Error (MSE) and the coefficient of determination (R^2), evaluated

on the test dataset. Lower MSE and higher R^2 indicate better predictive accuracy. To quantitatively evaluate the predictive performance of the machine learning models, we employ Mean Squared Error (MSE) and R^2 (coefficient of determination) as evaluation metrics. All models are trained to predict the log-transformed box office revenue, which helps reduce skewness and stabilize variance in the response variable. As shown in **Figure 3** the Random Forest model achieves the lowest MSE and the highest R^2, indicating superior predictive accuracy and a strong ability to explain variance in the target variable. This suggests that Random Forest effectively captures complex non-linear relationships between movie attributes and box office revenue. The Support Vector Machine (SVM) demonstrates competitive performance, with moderate MSE and relatively high R^2, indicating good generalization ability but slightly weaker explanatory power compared to Random Forest. In contrast, both the Generalized Additive Model (GAM) and Linear Regression exhibit higher MSE and lower R^2 values. While GAM allows for smooth non-linear effects, its performance remains limited by its additive structure. Linear Regression performs worst overall, reflecting its inability to capture complex non-linear interactions present in the data. Overall, these results justify the selection of Random Forest as the best-performing model for downstream scenario-based predictions.

Figure 4: Residual Density Comparison Across Models



This figure shows the kernel density distributions of prediction residuals (actual minus predicted values) on the log-transformed box office revenue for four models. Residuals closer to zero with narrower distributions indicate better model fit and lower prediction error. In addition to aggregate performance metrics, we further analyze model behavior using residual density distributions, as illustrated in Figure 4. Residuals are computed as the difference between the observed and predicted log box office revenue, allowing direct comparison across models. The Random Forest model exhibits the most concentrated residual distribution around zero, with a sharp and narrow peak. This indicates smaller prediction errors and greater consistency across observations, reinforcing its superior performance observed in MSE and R^2 metrics. The SVM model also shows a relatively narrow residual distribution, though with slightly heavier tails, suggesting occasional larger prediction errors. Nevertheless, its residual behavior remains more stable than that of GAM and Linear Regression. In contrast, GAM and Linear Regression display wider and flatter residual distributions, indicating greater variance in prediction errors. This

pattern suggests underfitting, as these models struggle to fully capture non-linear relationships and complex feature interactions present in the dataset. Residual density analysis complements quantitative metrics by providing insights into error structure and model robustness, further supporting the conclusion that Random Forest offers the most reliable predictive performance for this task.

4.2.5 AI Recipe Generator: Scenario-Based Content Generation Module

(1) Motivation and Role in the System

While the machine learning models focus on predicting box office revenue based on historical movie attributes, the AI Recipe Generator serves a complementary role by enabling scenario-based creative exploration. Rather than predicting numerical outcomes, this module provides text-based recommendations—referred to as movie recipes—that suggest how a movie could be designed or adjusted to meet specific strategic goals, such as maximizing revenue, improving ratings, or balancing both. Importantly, this module does not reuse the trained regression models described in the previous section. Instead, it operates as a generative AI layer that transforms user-defined constraints into creative suggestions.

(2) Core Idea of the AI Recipe Generator

The AI Recipe Generator is built on the idea that movie planning can be treated as a conditional text generation problem. Given: Investment goals (e.g., target box office revenue), Optimization strategy (revenue-driven, rating-driven, or balanced), Content constraints (genre, originality level), Creativity controls (temperature, sampling strategy), the system generates structured textual recommendations describing: Suggested genres and themes, Creative direction, Risk level and originality, Narrative or production style cues. This approach allows users to explore "what-if" creative scenarios that are difficult to capture using purely numerical prediction models.

(3) Content Constraints and User Inputs

The generator incorporates explicit constraints provided by the user, including: primary genre selection, desired originality level, target box office revenue (as a reference goal). These constraints act as conditioning signals, ensuring that generated recipes remain aligned with user expectations rather than producing unconstrained or irrelevant suggestions.

(4) Relationship to the Machine Learning Models

It is important to clarify that the AI Recipe Generator operates independently from the machine learning regression models used earlier in the project. Machine learning models → numerical prediction and evaluation, AI Recipe Generator → creative exploration and scenario simulation, However, the two components are conceptually connected: ML models provide quantitative

insights into what drives box office success. The AI generator translates these insights into qualitative, actionable ideas. Together, they form a hybrid decision-support system that combines data-driven prediction with creative AI generation.

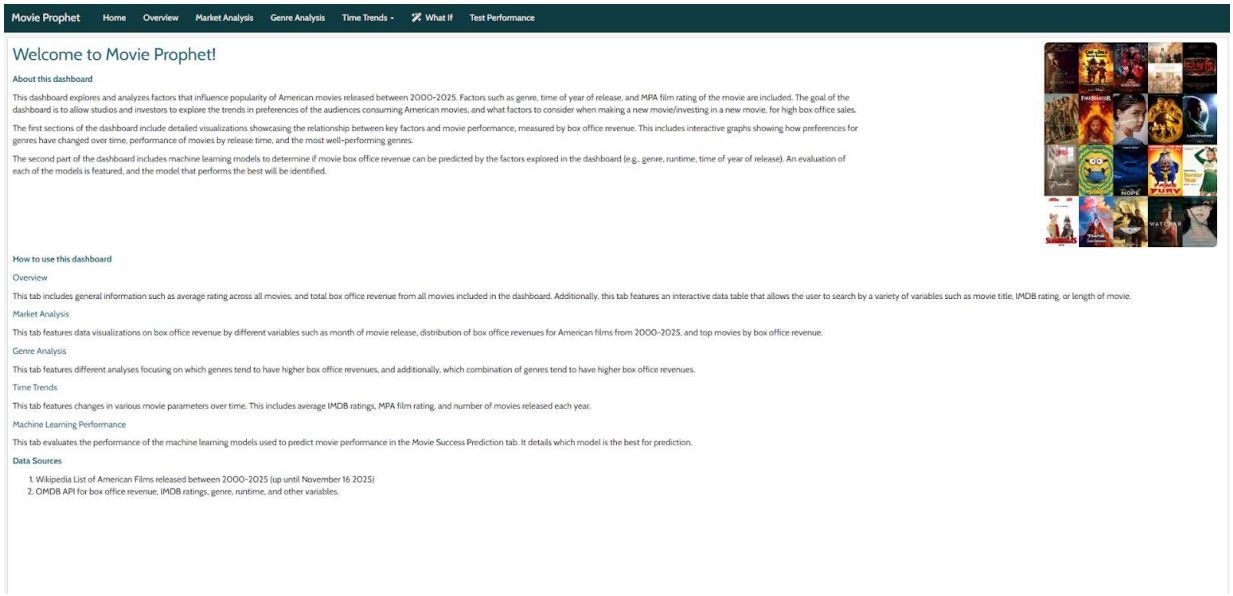
5. Data Analytic Product

5.1 Overview

The final data-analytic product is presented as a Shiny dashboard, which serves as a flexible and interactive platform. It provides comprehensive analytical insights and allows users to explore how various features—such as genre, rating, and more—relate to box office revenue for movies released between 2000 and 2025.

5.2 Home Page

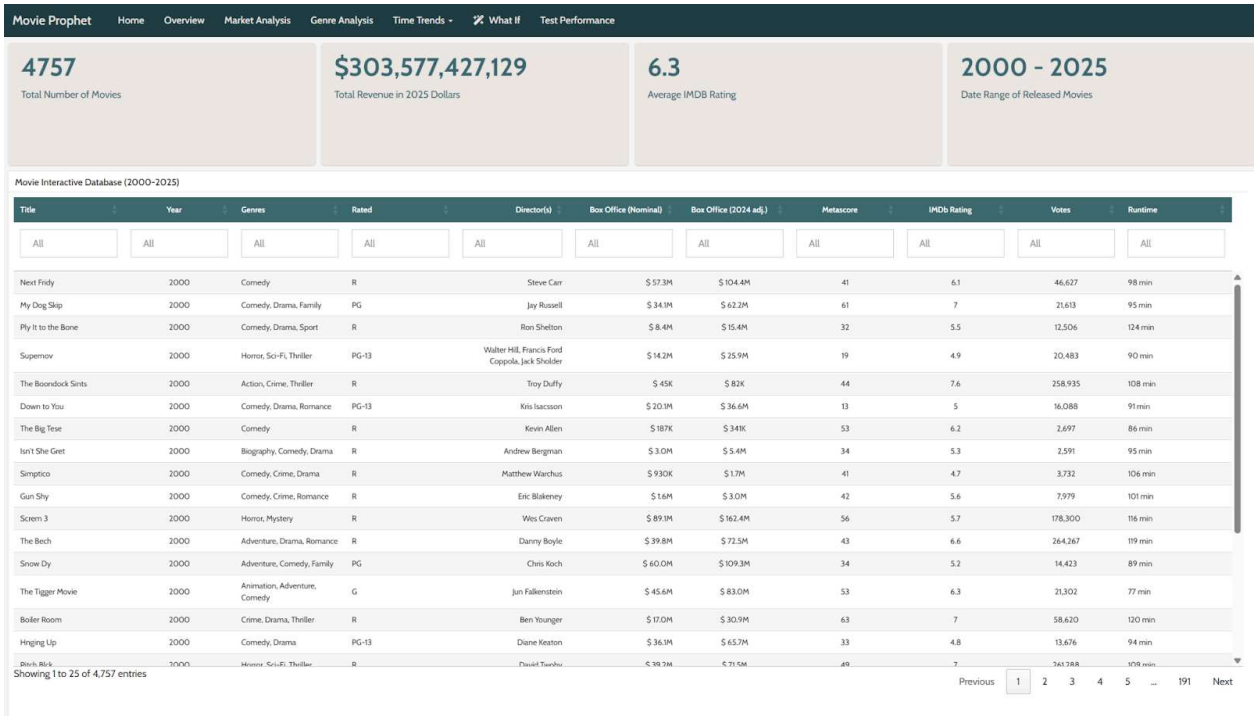
Figure 5: Home Page for shiny dashboard



The home page aims to introduce users to the purpose, scope, and structure of the dashboard, explaining how movie data from 2000–2025 is analyzed to understand factors influencing box office performance. It clearly outlines what insights users can gain—from trends in genres, release timing, and ratings to deeper market and genre analyses—and how these insights support data-driven decisions for studios and investors. The page also acts as a navigation guide, briefly describing the role of each tab (Overview, Market Analysis, Genre Analysis, Time Trends, and Machine Learning Performance) so users know where to explore specific questions. And it also includes a data source at the end for giving credits.

5.3 Overview Page

Figure 6: Overview Page for shiny dashboard



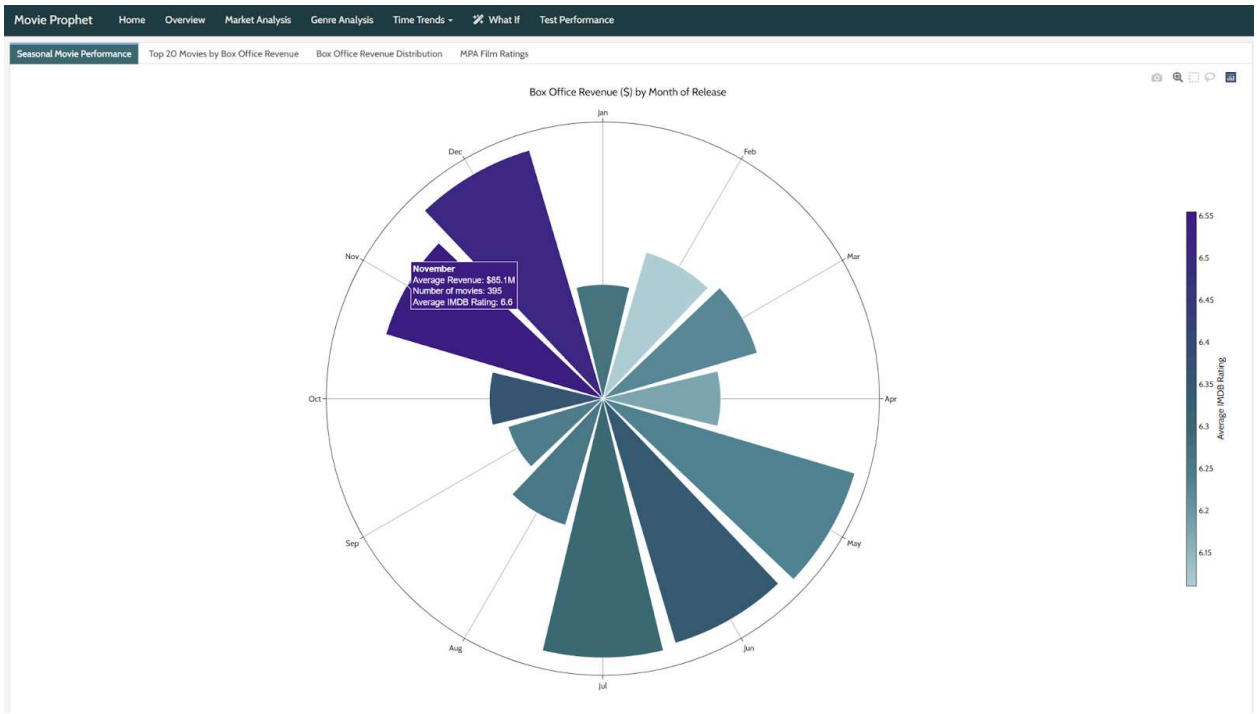
This page serves as the central overview and exploration hub of Movie Prophet, giving users a snapshot of the entire dataset while enabling deep, flexible exploration. At the top, key summary metrics highlight the scale and scope of the analysis, including the total number of films, cumulative box office revenue adjusted to 2025 dollars, average IMDb rating, and the release-year range covered. Below, the interactive movie database allows you to filter, search, and sort thousands of films by title, year, genre, rating, revenue, critical scores, and runtime, making it easy to uncover patterns or investigate individual titles in detail. Together, these elements are designed to let users quickly grasp the big picture and then seamlessly drill down into the specific movies and attributes that matter most to your questions.

5.4 Market analysis

The Market Analysis page brings together multiple complementary visualizations to help users to explore how the movie market behaves across timing, scale, and classification. By examining seasonal release patterns, top-grossing films by year, the overall distribution of box office revenue, and performance across MPA film ratings, this page connects high-level market structure with specific drivers of success. Together, these insights provide users with a comprehensive view of box office dynamics and the market forces shaping commercial movie performance.

5.4.1 Seasonal Movie Performance

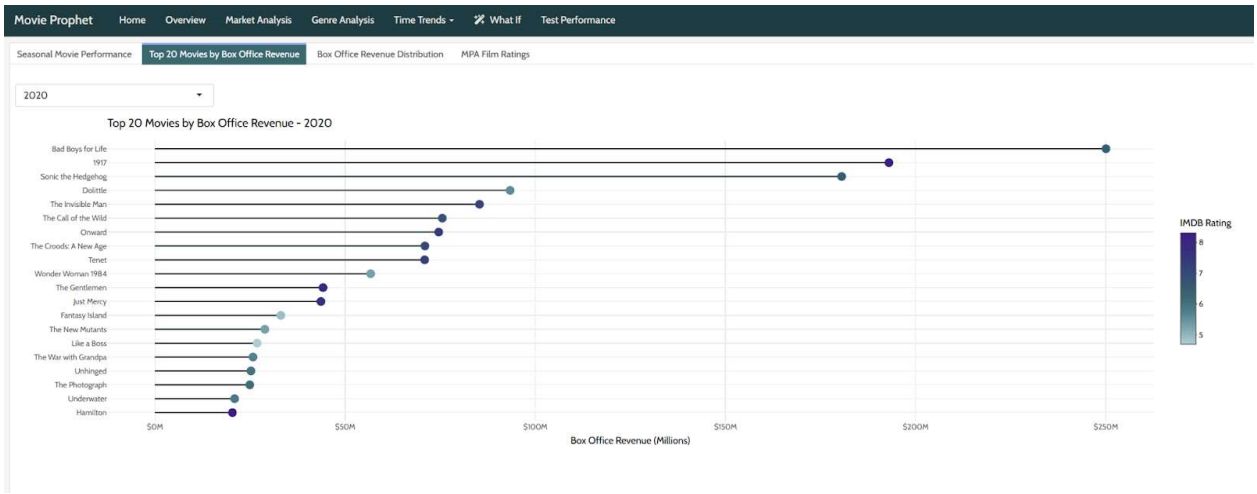
Figure 7: Sub Page "Seasonal Movie Performance" for Market Analysis Page



The first sub page functions as a visualization that illustrates seasonal movie performance by month of release, helping users understand how timing influences box office outcomes. Each radial bar represents a month, with its length corresponding to the average box office revenue generated by movies released during that period, allowing for quick comparison across the calendar year. Color intensity encodes the average IMDb rating, adding an additional layer of insight into how audience reception varies by release month. Interactive tooltips reveal key details—such as average revenue, number of releases, and average rating—so users can move seamlessly from high-level patterns to precise values. Together, these elements are designed to highlight seasonal trends, revealing when films tend to perform strongest financially and how critical reception aligns with those release windows.

5.4.2 Top 20 Movies By Box Office Revenue

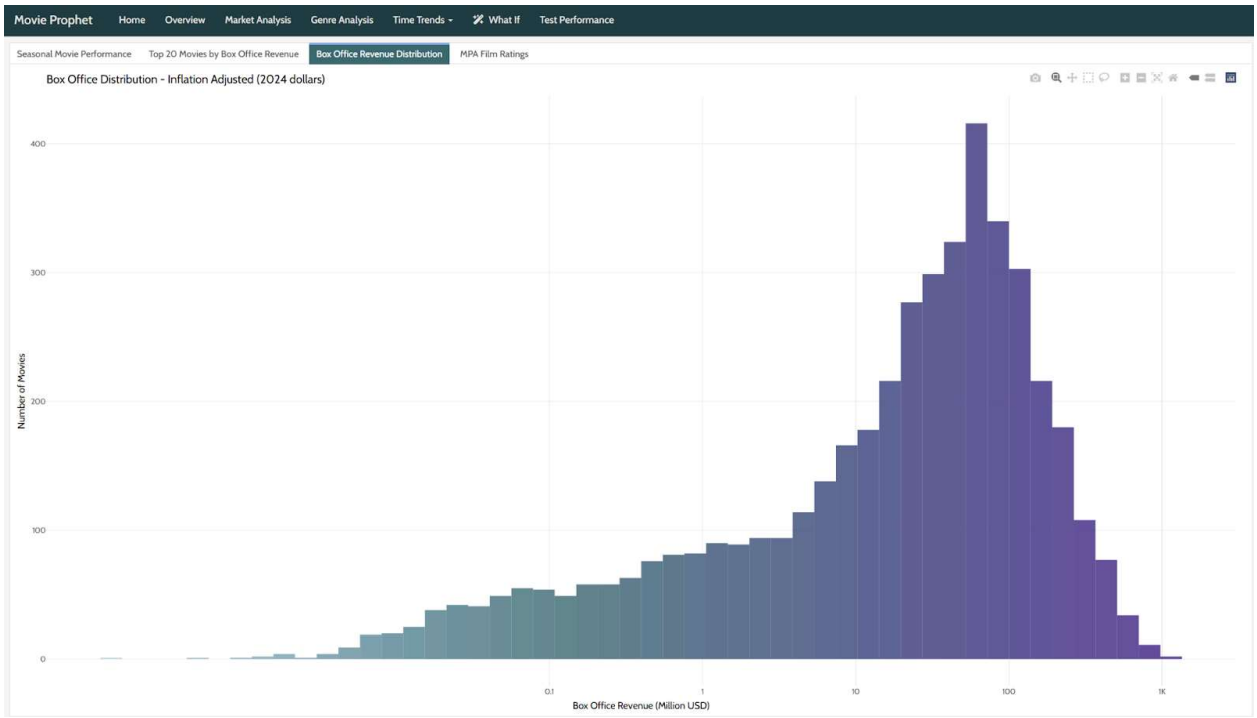
Figure 8: Sub Page " Top 20 Movies By Box Office Revenue" for Market Analysis Page



The second sub pages, top 20 movies by Box Office revenue, is an interactive chart that highlights the top 20 movies by Box Office revenue for a selected release year, allowing users to compare the highest-grossing films within a specific timeframe. Each horizontal bar represents a movie, with its length indicating total box office revenue in millions of dollars, while the color of each marker encodes the movie's IMDb rating to provide additional context on audience reception. The year selector enables users to scroll through different years and instantly update the chart, making it easy to explore how top-performing films change over time. Together, the ranked layout, color encoding, and interactive year control are designed to help users quickly identify standout box office successes and examine how commercial performance aligns with critical ratings across different release years.

5.4.3 Box Office Revenue Distribution

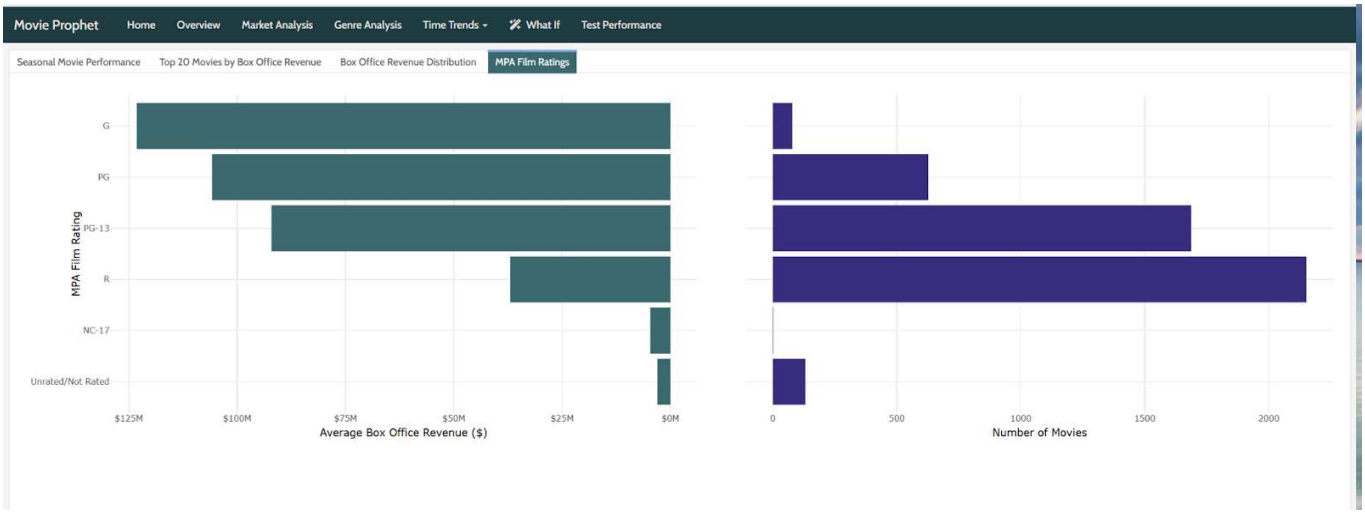
Figure 9: Sub Page "Box Office Revenue Distribution" for Market Analysis Page



The third sub pages, Box Office revenue distribution is a histogram visualizes the distribution of box office revenues across all movies, with figures adjusted for inflation to 2024 dollars to ensure fair comparison over time. The x-axis displays box office revenue on a logarithmic scale, capturing the wide range between low-earning films and major blockbusters, while the y-axis shows the number of movies within each revenue bin. Color gradients are used to reinforce changes in revenue magnitude, helping users quickly identify where films are most densely concentrated. By presenting the full revenue distribution in a single view, this graph enables users to understand the overall structure of the movie market, highlighting the imbalance between a small number of extremely high-grossing films and the large volume of lower-revenue releases.

5.4.4 MPA Film Ratings

Figure 10: Sub Page "MPA Film Ratings" for Market Analysis Page



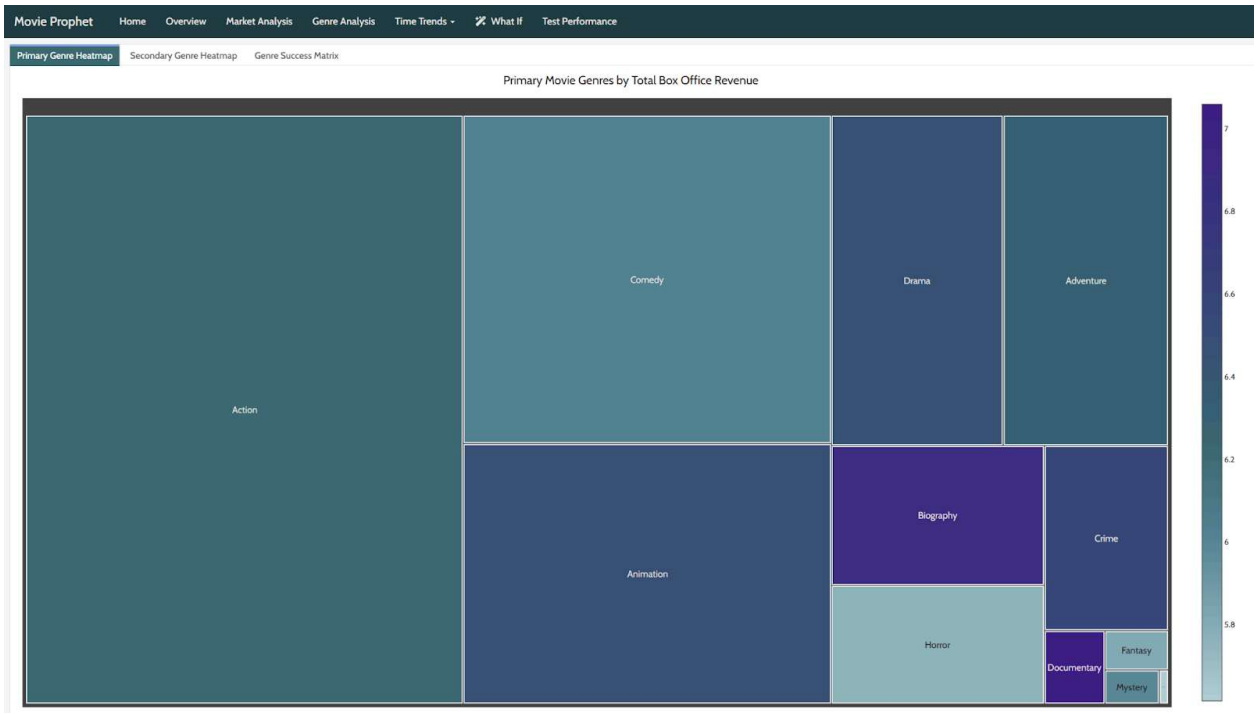
The last sub page, MPA film rating, is a paired bar chart that examines movie performance across MPA film ratings, combining financial outcomes with release volume to provide a balanced view of the market. The left panel displays the average box office revenue for each rating category, allowing users to compare how films rated G, PG, PG-13, R, NC-17, and Unrated perform on average at the box office. The right panel shows the number of movies released under each rating, highlighting how frequently each rating appears in the dataset. Together, these elements help users distinguish between ratings that generate higher average revenue and those that dominate in sheer volume, offering insight into the trade-off between commercial potential and production frequency across rating categories.

5.5 Genre Analysis

The Genre Analysis page is designed to help users understand how movie genres—and combinations of genres—shape box office performance. Through primary and secondary genre treemaps, users can quickly assess which genres contribute most to total revenue and how audience reception varies across them. The genre success matrix extends this view by examining interactions between primary and secondary genres, revealing which genre pairings tend to perform strongest in the market. Together, these visualizations provide users with a structured, comparative view of genre-driven dynamics, supporting deeper insight into how genre selection and combination influence commercial success.

5.5.1 Primary Genre Heatmap

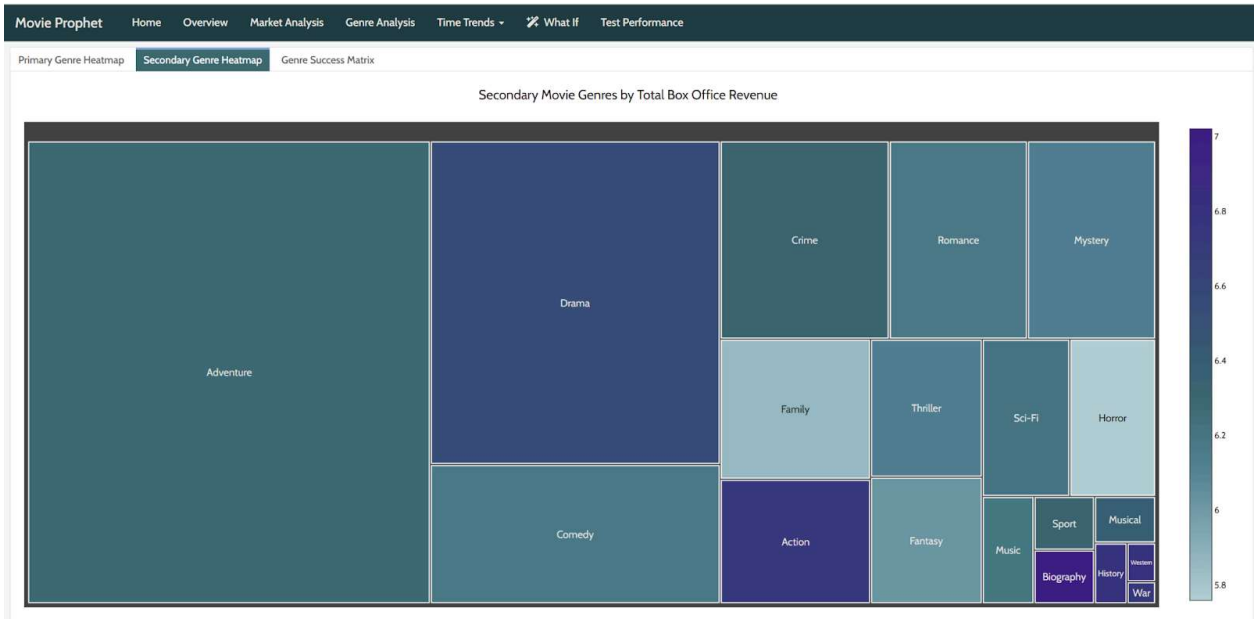
Figure 11: Sub Page "Primary Genre Heatmap" for Genre Analysis Page



The first sub page, primary genre heatmap, visualizes primary movie genres by total box office revenue, providing users with an immediate sense of how different genres contribute to overall market performance. Each rectangle represents a genre, with its size proportional to cumulative box office revenue, allowing users to quickly compare the relative financial impact of each genre. Color shading encodes the average IMDb rating, adding context on audience reception alongside commercial success. By combining scale and color in a single view, this graph enables users to identify genres that dominate revenue, those that balance strong earnings with higher ratings, and smaller niches with distinct performance profiles, offering a clear and intuitive overview of genre-level market dynamics.

5.5.2 Secondary Genre Heatmap

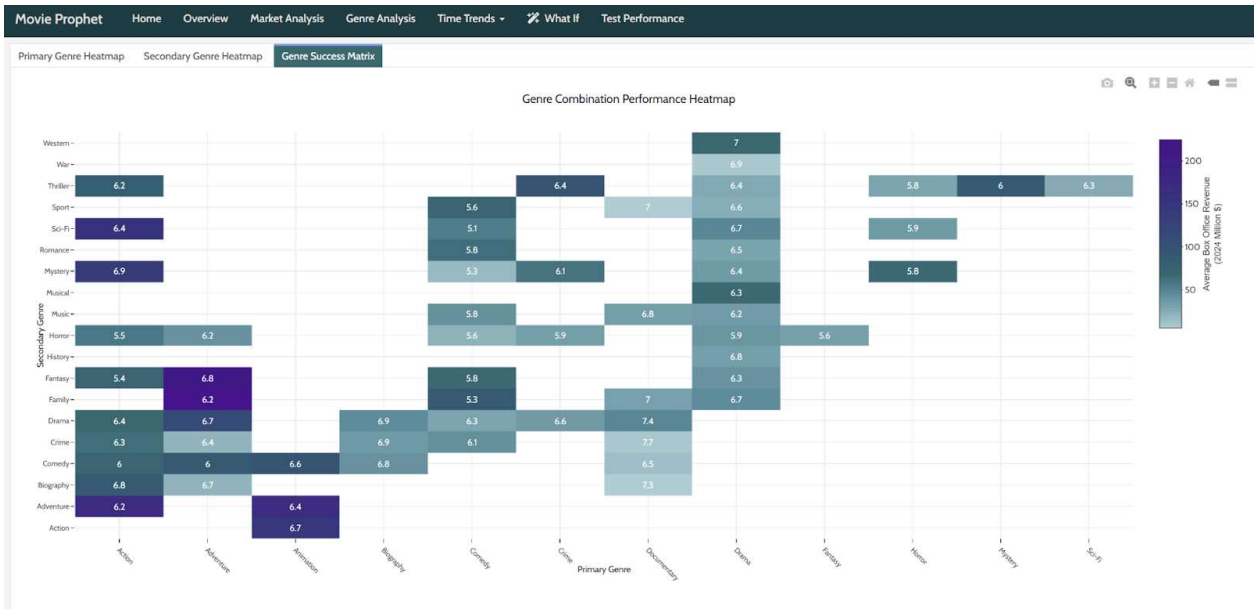
Figure 12: Sub Page "Secondary Genre Heatmap" for Genre Analysis Page



The second sub page, secondary genre heatmap provides a concise overview of secondary movie genres by total box office revenue, offering users a complementary perspective to the primary genre analysis. The heatmap is structured in the same way as the first sub page, primary genre heatmap. The only difference is it aims to reflect the secondary genre contribution instead of the primary genre.

5.5.3 Genre Success Matrix

Figure 13: Sub Page "Genre Success Matrix" for Genre Analysis Page



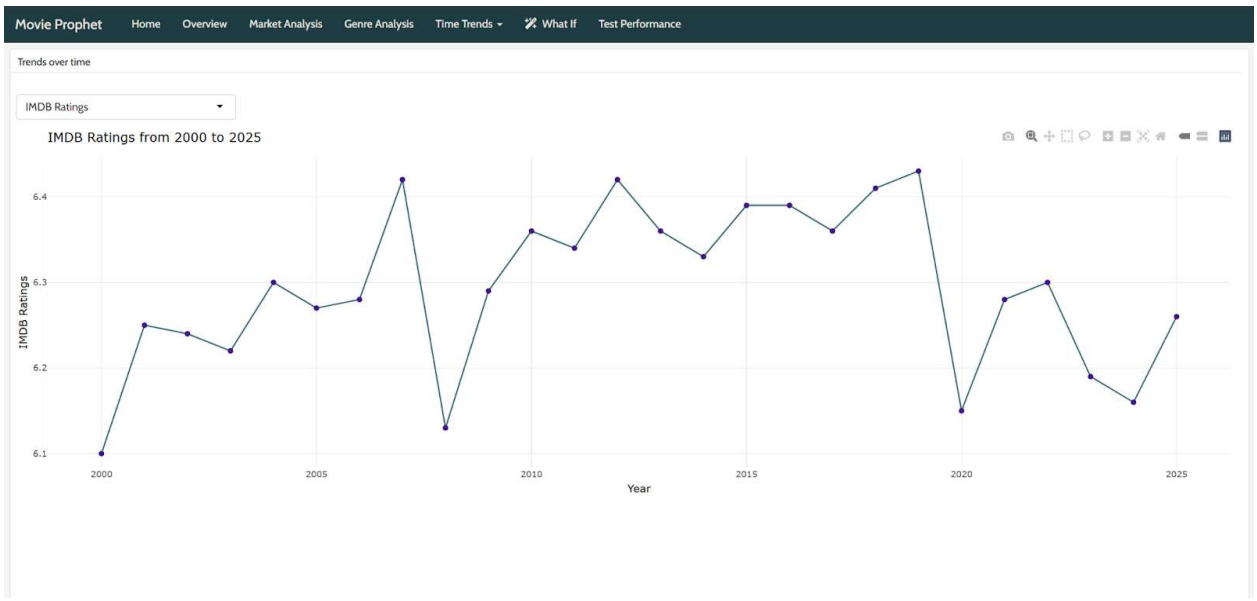
The third sub page, genre success matrix, aims to explore performance across combinations of primary and secondary genres, allowing users to examine how genre pairings influence box office outcomes. The x-axis lists primary genres and the y-axis lists secondary genres, while each cell represents the average box office revenue (inflation-adjusted to 2024 dollars) for movies that share that specific genre combination. Color intensity reflects revenue magnitude, and embedded labels provide quick reference values for comparison. By visualizing genre interactions in this matrix format, the chart enables users to identify which genre pairings tend to perform strongly, which combinations are less successful, and how blending genres shapes overall market performance.

5.6 Time trends

The Time Trends page is designed to help users explore how key movie characteristics and industry patterns evolve over time. By combining trend lines for quantitative metrics such as IMDb ratings, runtime, and box office revenue with categorical views showing changes in genre composition, this page offers both depth and flexibility in temporal analysis. Interactive controls allow users to select the parameters they are most interested in, enabling direct comparison of different metrics and classifications across years. Together, these visualizations provide users with a clear, longitudinal perspective on how audience reception, production attributes, and genre emphasis have shifted from 2000 to 2025.

5.6.1 Numerical Parameter Plot

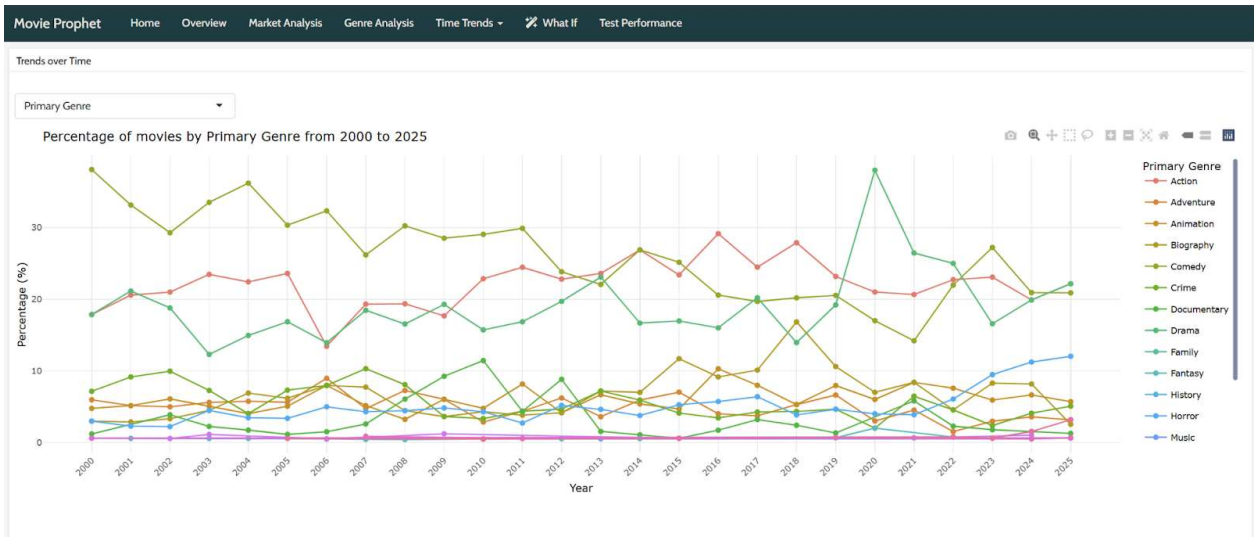
Figure 14: First Plot "Numerical Parameter Plot" for Time Trends Page



The first plot is an interactive time-series chart that tracks changes in key movie metrics over time, allowing users to explore how industry characteristics evolve from 2000 to 2025. The line and markers represent yearly averages for the selected parameter—such as IMDb ratings, runtime, or box office revenue—with the x-axis showing release year and the y-axis reflecting the chosen metric. A dropdown menu enables users to switch seamlessly between different parameters, instantly updating the visualization to reflect their area of interest. By combining clear trend lines with interactive controls, this graph helps users identify long-term patterns, shifts, and anomalies in movie performance and production characteristics across time.

5.6.2 Categorical Parameter Plot

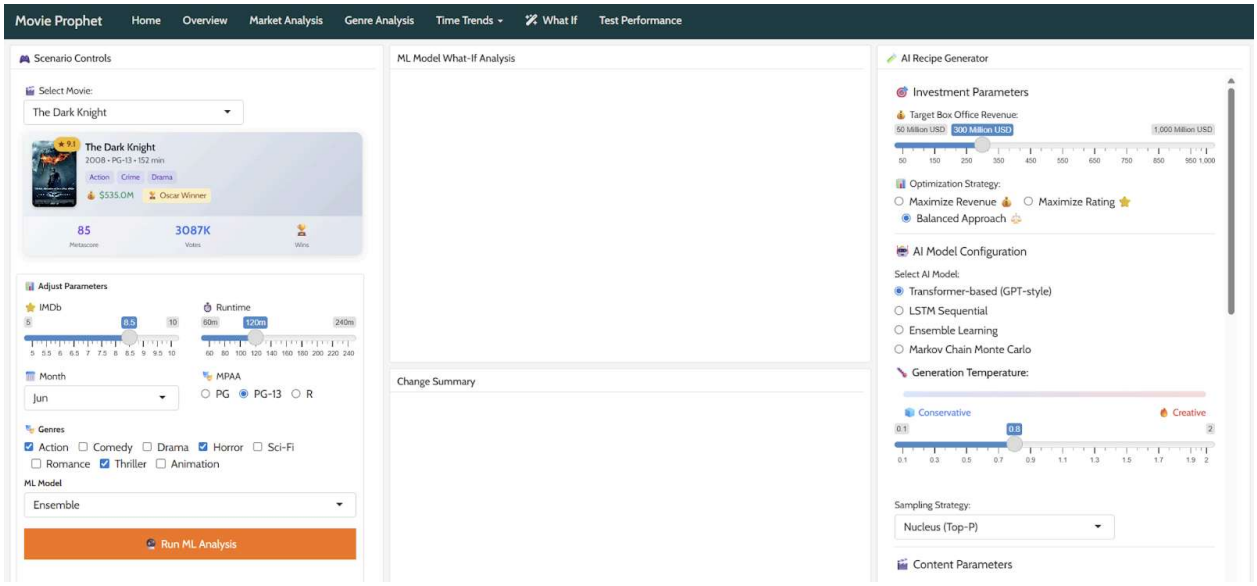
Figure 15: First Plot "Categorical Parameter Plot" for Time Trends Page



The second plot is an interactive time-trend visualization that shows how the composition of movies changes over time by tracking the percentage of films across selected categorical parameters from 2000 to 2025. Each line represents a category—such as a primary genre in the view shown—with the x-axis indicating release year and the y-axis showing the proportion of movies released in that category. An interactive dropdown menu allows users to switch between different parameters, including primary genre, secondary genre, and other categorical groupings, instantly updating the chart to reflect the selected perspective. By combining multi-line trends with flexible parameter selection, this graph enables users to explore shifts in industry focus, emerging patterns, and long-term changes in how movies are categorized over time.

5.7 What If

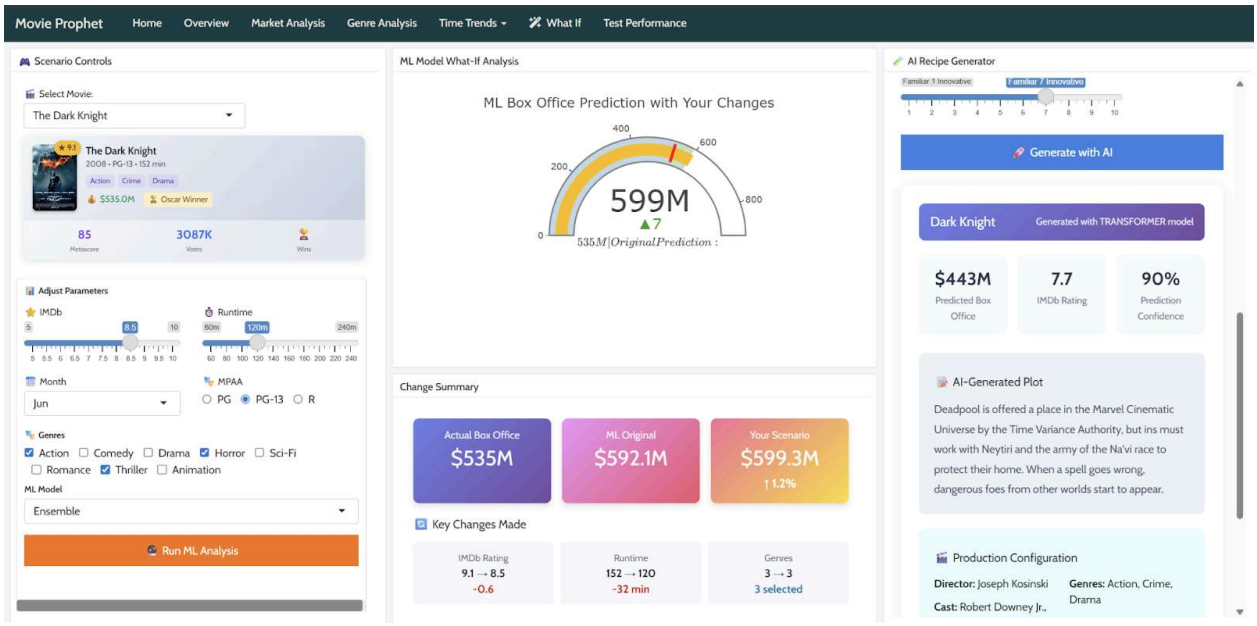
Figure 16: What If Page for the shiny dashboard



The What-If Analysis page allows users to simulate alternative movie scenarios and explore how changes in key attributes may influence predicted performance. Users begin by selecting an existing movie as a baseline, then adjust parameters such as IMDb rating, runtime, release month, MPA rating, and genre composition using intuitive sliders, dropdowns, and checkboxes. Once configured, users can run the machine-learning analysis to see how these changes affect projected outcomes, with results summarized in the central analysis and change summary panels. It enables investors to define investment goals, choose optimization strategies.

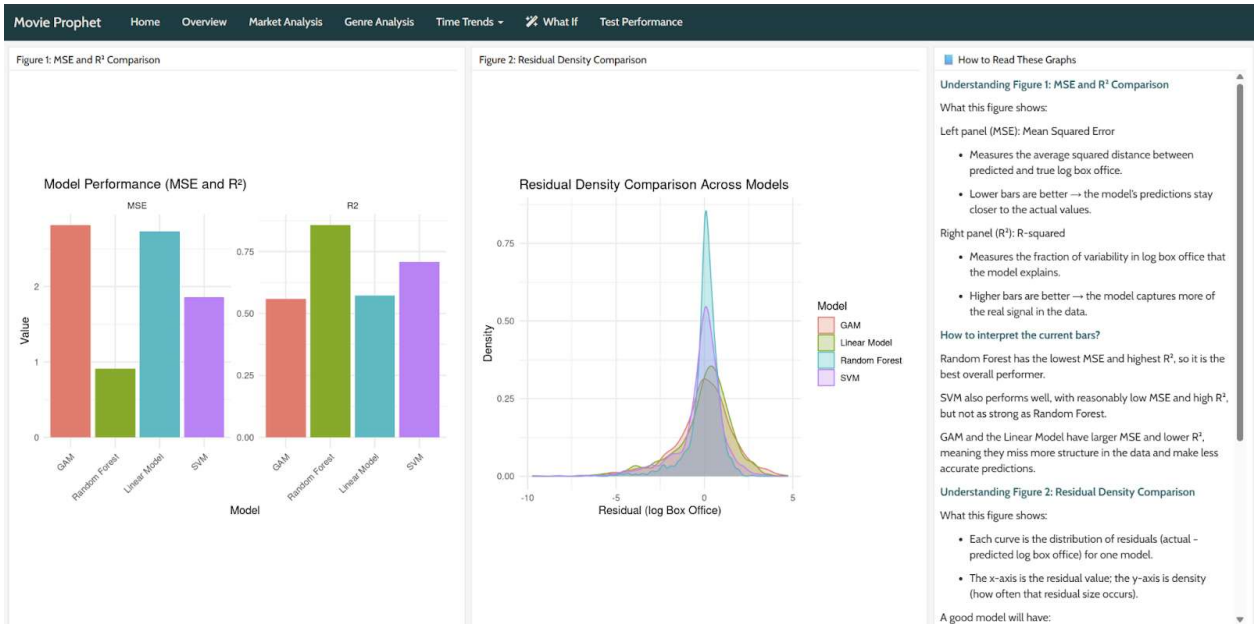
On the right, the AI Recipe Generator serves as a creative companion to the What-If Analysis, allowing users to translate data-driven decisions into narrative concepts. Users begin by selecting content parameters such as the primary genre and desired level of originality, balancing familiar elements with innovation using an intuitive slider. With a single action, the system generates an AI-crafted movie concept alongside predicted outcomes, including estimated box office revenue, projected IMDb rating, and confidence in the prediction. By pairing creative plot generation with quantitative performance signals, this tool helps users explore how different storytelling choices may align with market potential, bridging analytical insight and creative experimentation in a single, streamlined workflow.

Figure 17: Example Output for What If Page(including what-if analysis and recipe generator)



5.8 Test Performance

Figure 18: Test Performance Page for the shiny dashboard



The Test Performance page is designed to transparently evaluate and communicate how different machine-learning models perform when predicting box office outcomes, with a focus on accessibility for non-technical users. The visualizations compare models using intuitive performance metrics—such as prediction error and explanatory power—while the accompanying residual analysis illustrates how closely each model's predictions align with actual results. To support investors who may not have a background in machine learning, a dedicated explanation panel on the right provides clear, plain-language guidance on how to read and interpret each chart and what the results imply. Together, this page helps users build confidence in the predictive models by combining rigorous evaluation with clear interpretation, ensuring that model insights are both credible and understandable for decision-making.

6. Conclusion

6.1 Results

As the producer of this platform, the goal of these visualizations is to help users move beyond isolated metrics and instead develop a coherent, market-level understanding of how films perform, why they succeed, and where risk and opportunity tend to concentrate. Taken together, the charts reveal that box office outcomes are not random, but shaped by a combination of timing, genre positioning, audience structure, and long-term industry trends. Rather than treating each film as a standalone bet, users are encouraged to view projects within a broader economic and historical context.

The seasonal performance chart demonstrates that certain months(those in winter and summer) consistently deliver stronger average revenues, often accompanied by solid audience reception. This suggests that demand cycles, holiday effects, and competitive spacing materially influence outcomes. For users evaluating investment opportunities, timing insights help distinguish between underperformance caused by weak fundamentals and underperformance driven by unfavorable release windows. In many cases, optimizing timing can meaningfully improve return potential without altering creative scope.

At the same time, the box office distribution view underscores a structural reality of the film industry: revenues are highly skewed. Most titles cluster in modest performance ranges, while a small number of films generate disproportionate returns. By visualizing this distribution on a log scale, users can clearly see the asymmetry that defines film investing. This perspective is critical for setting realistic expectations and reinforces the importance of portfolio thinking—accepting that many projects will be moderate performers while a few successes carry the bulk of value creation.

Genre-focused visualizations deepen this analysis by revealing how revenue concentration and audience reception vary across both primary and secondary genres. Treemap views make it immediately clear which genres act as consistent revenue anchors and which occupy more specialized or quality-driven roles. Some genres function as consistent revenue anchors both as primary and secondary genres, such as comedy, while others serve as strong anchors only when positioned as the primary genre, such as action. In contrast, certain genres—such as fantasy—tend to generate lower average revenue whether they appear as a primary or secondary genre. These patterns indicate that audience demand varies substantially across genres, and that not all genres carry the same level of commercial appeal in the market. Meanwhile, these charts show that large market share does not always coincide with the highest audience ratings, helping users distinguish between scale-driven returns and reputation-oriented performance. This distinction is particularly useful when evaluating whether a project is designed to maximize reach, prestige, or a balance of both. The genre success matrix extends this logic by highlighting the performance of genre combinations rather than single categories in isolation. Certain pairings consistently outperform, suggesting that complementary genre blending can broaden audience appeal or improve market fit. For users assessing early-stage concepts, these insights help identify combinations that historically align with stronger outcomes and avoid pairings that may limit commercial traction.

In combination, these visualizations transform complex market data into an integrated decision framework. They enable users to evaluate projects through multiple lenses—timing, positioning, scale, risk, and trajectory—while maintaining clarity and accessibility. For an investor-focused audience, this holistic view supports more disciplined reasoning, stronger justification of

assumptions, and ultimately, more confident investment decisions grounded in evidence rather than intuition alone.

6.2 Limitations & Bias

There are several limitations for our project and findings. First, the dataset is limited to American films released between the years 2000 and 2025. This is a relatively small number of movies compared to the number of movies produced globally and in a variety of languages. Additionally, some movies with our desired criteria were not able to be included because they were either not in the OMDB API, or did not have box office revenue information. Further, during data collection we obtained data from the OMDB API by title of movies. However, it would have been more accurate to use the unique IMDB ID, particularly if movies have the same titles.

Beyond data availability and collection issues, the machine learning and AI components of this project also introduce several limitations. First, the predictive models are built on historical features and trends, which may limit their ability to capture novel or disruptive factors that influence future box office performance. As a result, predictions may favor patterns commonly observed in past films. Second, box office revenue shows high variance and extreme outliers, which can bias model training toward blockbuster movies and reduce prediction accuracy for low- or mid-budget films, even after feature engineering and data transformation. Finally, while more complex models such as Random Forest and SVM improve predictive accuracy, they reduce interpretability compared to simpler models.

6.3 Future Work

Future work on this project can expand its impact by addressing current limitations in data scope, market coverage, and analytical depth. One natural extension is to broaden the dataset, as the movies analyzed here represent only a subset of the overall film landscape. Larger and more comprehensive datasets could be incorporated, such as films listed on Wikipedia from 1950 to 2025, enabling long-horizon trend analysis across multiple cinematic eras. Alternatively, a dataset covering all Wikipedia-listed films from 2000 to 2025—including smaller and less commercially prominent titles—could provide a more representative view of the modern film ecosystem. In addition, this project focuses exclusively on movies released in the American market, which constrains the generalizability of the conclusions. Incorporating data from other regions would allow for comparative analysis across global markets and reveal how genre preferences, seasonal effects, and revenue dynamics differ internationally.

On the machine learning side, the project currently relies on a set of commonly used baseline models; future work could explore more advanced or specialized approaches and systematically evaluate whether they yield stronger predictive performance. Finally, while this project emphasizes visual exploration to illustrate how various factors relate to box office revenue,

complementary factor-weight or feature-importance analyses could be introduced. Such methods would provide a more quantitative understanding of how individual variables contribute to revenue outcomes, offering an additional layer of insight beyond descriptive visualization.

7. Data Sources

I. Wikipedia List of American Films 2000 - 2025 (e.g., https://en.wikipedia.org/wiki/List_of_American_films_of_2000)
II. OMDB API (https://www.omdbapi.com/)
III. priceR R package (https://cran.r-project.org/web/packages/priceR/refman/priceR.html#adjust_for_inflation)

8. Appendix

Link to Data Dashboard:
https://rlim8.shinyapps.io/MovieProphet/#section-home