MOVIE REVENUE PREDICTION IN R

CMIS 566 Introduction to Business Analytics

Group 7
Hannah Smith
Cristopher Isada
Sumir Sharma Acharya
Craig Schafer
Submitted to: Dr. Prajakta Kolte

Contents

List of Figures

List of Tables

Introduction & Research Questions

Movies are a huge part of entertainment, but figuring out which ones will do well at the box office isn't easy. In our project, we took a closer look at film metadata and historical box office data from 2000 to 2024. We used statistical analysis and predictive modeling to dig into what might drive a film's financial success.

We centered our research on four main questions:

1. Which genres tend to be the most profitable?

2. How do elements like the director, cast, and IMDb ratings impact revenue?

3. Is it possible to predict a movie's revenue before it hits theaters?

4. Do higher IMDb ratings really correlate with increased revenue?

Through this research, we hope to better understand filmmaking as a business and show how data science can provide valuable insights into creative industries.

Approach and Potential Resources

We started our project by exploring datasets on platforms like Kaggle, but many were missing reliable metadata or financial details. To tackle this, we gathered raw data from IMDb, which included movie titles, release years, ratings, and more, and combined it with the "Movies Box Office Collection Data 2000–2024" from Kaggle for financial insights. Our approach involved cleaning and merging the datasets to prepare them for analysis. We applied exploratory data analysis techniques and built regression and classification models, all while using R programming for data manipulation and visualization throughout the project.

Citations

- Parth. "Movies Box Office Collection Data 2000–2024." Kaggle, 15 Aug. 2024. https://www.kaggle.com/datasets/parthdande/movies-box-office-collection-data-2000-2024
- IMDb. "IMDb Data Files Available for Download." Accessed June 5, 2025. https://datasets.imdbws.com/

Dataset Description

The final dataset used in our project combines structured movie metadata from IMDb with financial revenue data from a global box office collection file. The merged dataset includes over 2,000 films released between 2000 to 2024, filtered to include only English-language movies with calid worldwide figures.

Key variables in the dataset include:

Table 1: Description of Variables in the Dataset

Variable	**Measurement Scale**	**Description**	**Source**
title, year	Nominal	Movie name and release year	IMDb
genres	Nominal	One or more genres (e.g., Action, Comedy)	IMDb
avgRating	Interval	Average IMDb user rating	IMDb Ratings
directedBy	Nominal	Director's full name	IMDb Crew
Starring	Nominal	Top 3 billed cast members	IMDb Principals
domestic, foreign, worldwide	Ratio	Revenue breakdown by region	Box Office

Data Preparation and Cleaning

Load and Filter IMDb Metadata (2000–2024)

We started by loading the raw .tsv files from IMDb using appropriate R functions such as read_tsv () and filtered the title.basics.tsv file to include only entries where titleType == "movie" and startYear was between 2000 and 2024. To ensure consistency in language, we also filtered title.akas.tsv to include only English-language titles (language == 'en').

Handle Missing Values and Correct Formats

We identified and handled missing data in several key fields.

- Missing IMDb ratings were filled with the median rating across the dataset.
- Empty director fields were labeled as "Unknown".
- Revenue columns (e.g., Worldwide, Domestic) were originally stored as strings and formatted with commas, so we removed formatting and converted them to numeric.

Remove Duplicates

Before merging, we removed any duplicate entries based on a composite key of title and release year to avoid redundancy during join operations.

Enrich Metadata with IMDb Fields

To enhance the dataset, we performed joins using IMDb IDs across the following:

- title.ratings.tsv: Provided avgRating
- title.crew.tsv and name.basics.tsv: Provided full director names
- title.principals.tsv: Used to extract the top 3 credited actors for each movie, creating a starPower indicator

Standardized Titles for Dataset Merging

To ensure a reliable join between the IMDb dataset and the box office Excel file (which often use different naming conventions), we created a custom join_key. This was done by:

- Converting titles to lowercase
- Removing punctuation and extra whitespace
- Concatenating the cleaned title with the release year

This normalized key allowed for fuzzy matching across both datasets, even when formatting varied.

Merge with Box Office Revenue Data

Using the join keys, we performed a left join to merge box office revenue fields:

- Domestic
- Foreign
- Worldwide

We then filtered out:

- Any unmatched IMDb entries (i.e., with no valid revenue data)
- Any movies missing a valid Worldwide gross figure

Outlier Removal

To avoid distortions in revenue-based modeling, we removed outliers in the Worldwide revenue column using the 1st and 99th percentiles. This was done using quantile() in R and replacing extreme values with NA.

Feature Engineering

To enrich the dataset and support modeling:

- We extracted the primary genre from the genre list. (e.g, Action, Comedy).
- A new binary variable `recent` was created based on whether movie was released after the dataset's median year.
- We also created a `blockbusterDirector` variable, which identifies whether the movie's director is among the top 5 highest median-grossing directors.

This resulted in a clean, structured dataset that included only financially tracked films with metadata verified from IMDb.

Exploratory Data Analysis

Once we had a clean and enriched dataset, we moved into exploratory data analysis to better understand the relationship between features like genre, ratings, director reputation, and financial performance. This step helped us determine what features might be useful in predicting box office revenue and guided our model development.

Descriptive Statistics

- Worldwide revenue ranged from a few hundred thousand to multiple billion dollars (USD).
- The median worldwide gross was approximately $50 million, with a highly skewed distribution.
- Ratings were normally distributed around 6.5 – 7.0 out of 10.
- Most movies had between 2-3 top-billed actors, and a small number were directed by “blockbuster” director.

Revenue Distribution

We visualized revenue both in raw and log-transformed form to better understand spread and detect skew.

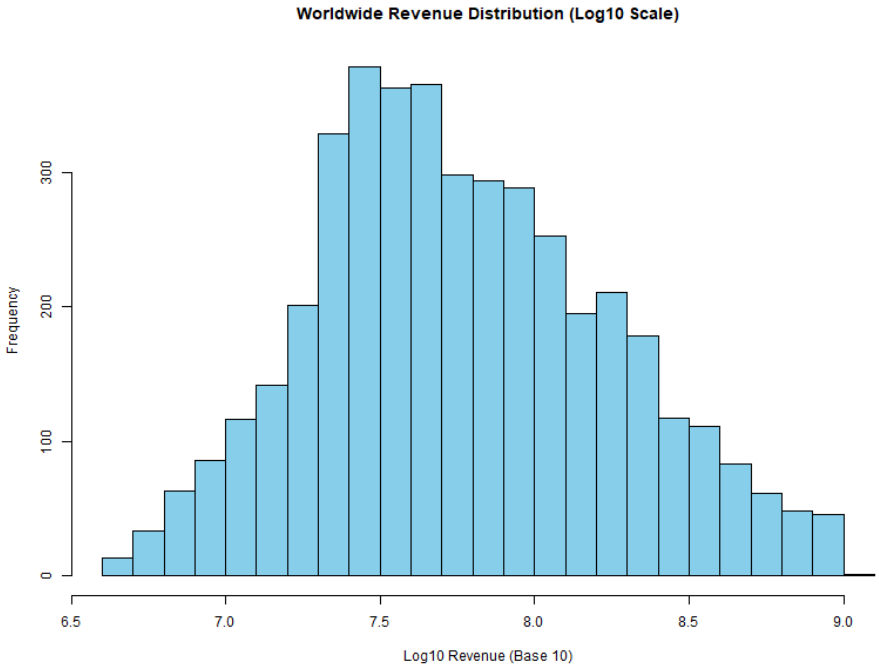

Figure 1: Histogram of Revenue (log10)

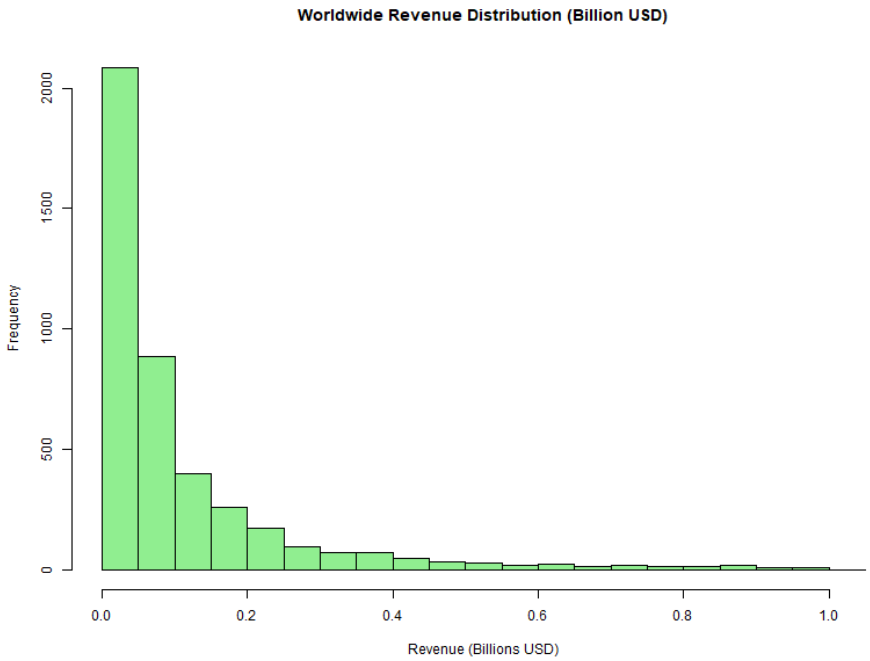

Figure 2: Histogram of revenue (Billions USD)

Genre-Based Analysis

We grouped movies by their `mainGenre` (first listed genre) to identify which categories are most profitable.

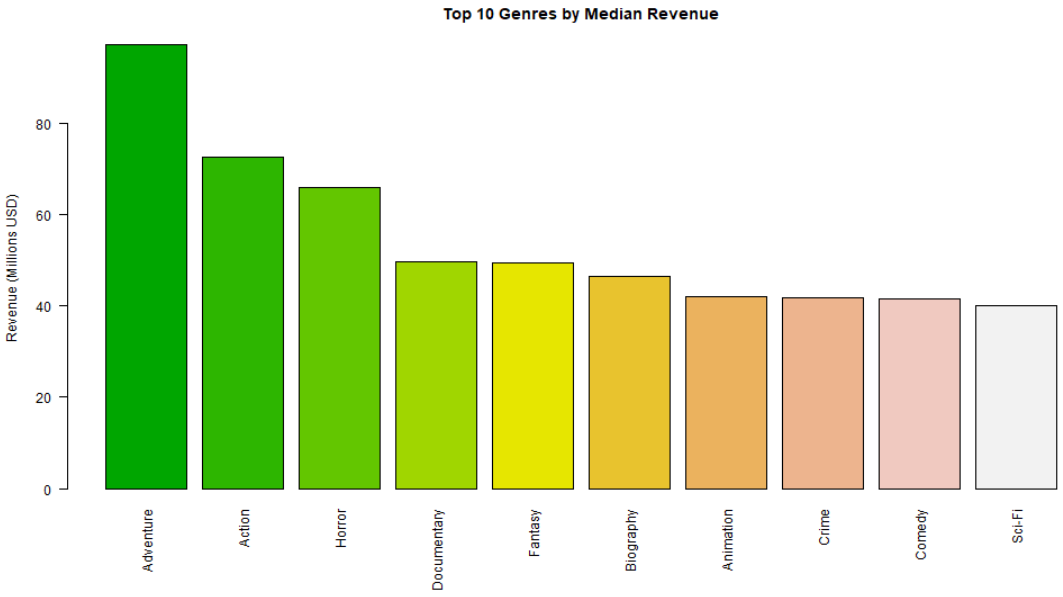

Figure 3: Bar chart of Top 10 genres by median revenue

Some important insights we gained from the chart were:

- Action, Adventure and Fantasy were consistently top-performing genres.
- Drama has the most entries but lower average revenue.

IMDb Rating vs Revenue

To investigate whether quality (as perceived by users) impacts revenue, we plotted IMDb ratings against log-transformed revenue.

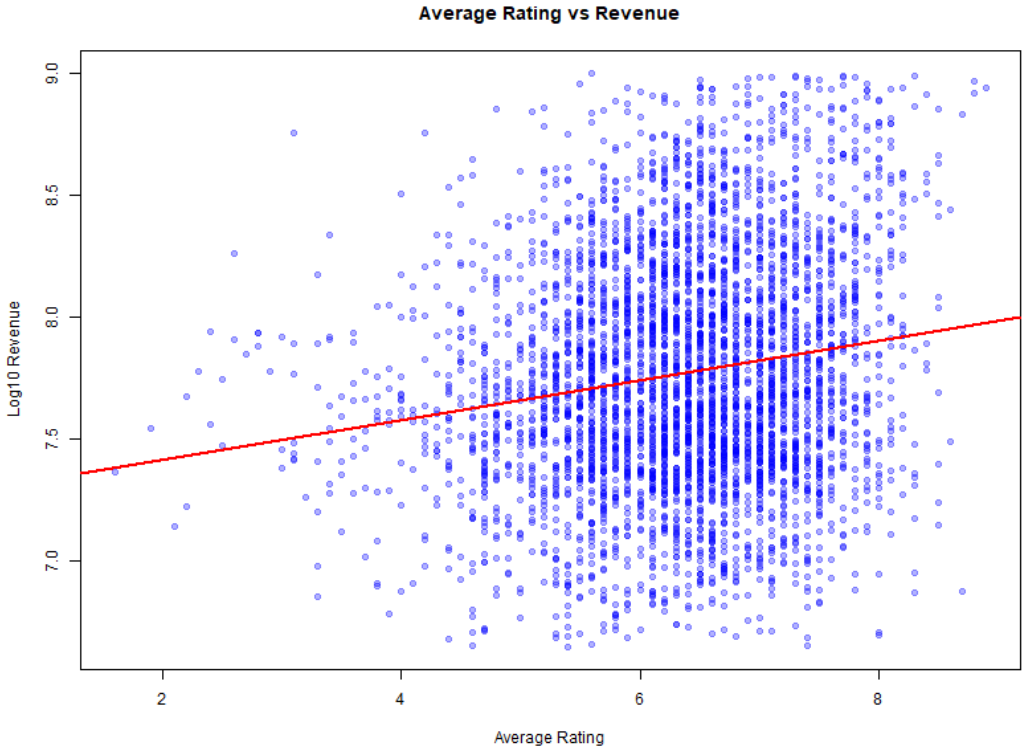

Figure 4: Scatter Plot with regression line

There is a mild positive correlation between higher ratings and higher revenue. Some highly rated movies still underperformed, indicating quality alone isn't enough.

Handling Outliers

To ensure model reliability, we removed movies in the bottom and top 1% of revenue using quartile thresholds.

Feature Engineering Overview

We created the following new features to enrich the modeling phase:

Table 2: Features Overview

Feature	Description
mainGenre	First listed genre from genre string
starPower	Number of top-billed cast members
recent	1 if movie released after median year
blockbusterDirector	1 if directed by a top 5 director by median revenue

Data Modeling

To understand and predict box office revenue, we built and tested several machine learning models. Two variations of our dataset were used: a full version with all available predictors (Model 1), and a reduced version (Model 2) excluding IMDb rating and recency. This allowed us to explore model performance with both detailed and minimal feature sets.

Each model was trained using a 70/30 train-test split. We evaluated performance using:

- RMSE and R^2 for regression tasks
- Accuracy and AUC for classification tasks

Logistic Regression

We used logistic regression to classify whether a movie would be a "hit" (when worldwide revenue of movie is over $100 Million. The model outputs a probability, and we classified hits when predicted probability > 0.5. This method helps understand which factors most influence a movie's chances of becoming successful.

Here's a table showing the dependent variable and independent variables used in modeling for Logistic Regression.

Table 3: Logistic Regression - List of Variables

Logistic Regression		
Models	Dependent (Target) Variable	Independent (Predictor) Variables
Model 1	hit	avgRating, starPower, mainGenre, recent, blockbusterDirector
Model 2	hit	starPower, mainGenre, blockbusterDirector

Decision Tree

Decision trees helped visualize the relationships between variables like genre, starPower, and rating and how they split to predict revenue. It's an interpretable model that gives clear insights like showing whether directors or genres dominate earnings.

Application in Project

By using these models, we aim to:

- Predict log-transformed worldwide revenue
- Classify whether a movie becomes a hit
- Compare model accuaracy
- Identify which variables influence results most

This model allows practical use cases like pre-release forecasting for new titles or understanding what kinds of films are statistically more likely to succeed.

Logistic Regression – Model 1

The glm() function in R was used to train the logistic regression model. Significant predictors included avgRating and blockbusterDirector ($p < 0.001$), suggesting both have strong

influence on hit probability. StarPower was not significant, indicating that a larger cast does not guarantee success. Residual deviance was lower than null deviance, showing good model fit.

```
> summary(logit_full)

Call:
glm(formula = hit ~ avgRating + starPower + mainGenre + recent +
    blockbusterDirector, family = "binomial", data = train)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.76525    1.40427  -1.257 0.208733
avgRating            0.44508    0.04886   9.110  < 2e-16 ***
starPower           -0.48474    0.45844  -1.057 0.290339
mainGenreAdventure   0.47321    0.13187   3.589 0.000333 ***
mainGenreAnimation  -0.46015    0.38328  -1.201 0.229928
mainGenreBiography  -1.03629    0.20354  -5.091 3.56e-07 ***
mainGenreComedy     -0.74185    0.11408  -6.503 7.89e-11 ***
mainGenreCrime      -1.12232    0.21623  -5.190 2.10e-07 ***
mainGenreDocumentary -1.43170    0.86920  -1.647 0.099527 .
mainGenreDrama      -1.21344    0.14061  -8.630  < 2e-16 ***
mainGenreFantasy    -0.77583    0.65073  -1.192 0.233167
mainGenreHorror     -0.15628    0.20966  -0.745 0.456040
mainGenreMystery    -0.98807    0.78940  -1.252 0.210689
mainGenreRomance    -1.16925    1.07523  -1.087 0.276841
mainGenreSci-Fi      0.06987    1.29500   0.054 0.956972
mainGenreThriller  -12.89339  235.56038  -0.055 0.956350
recent              -0.10224    0.08467  -1.207 0.227277
blockbusterDirector  2.15154    1.10604   1.945 0.051743 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3681.1  on 2991  degrees of freedom
Residual deviance: 3414.3  on 2974  degrees of freedom
AIC: 3450.3

Number of Fisher Scoring iterations: 12
```

Figure 5: Summary Output for Logistic Model 1

Confusion Matrix

A confsuion matrix was generated for Model 1 (Figure 6), which shows the following calssification results:

Table 4: Confusion Matrix for Model 1

	Actual: Not Hit (0)	**Actual: Hit (1)**
Predicted:0	837	315
Predicted:1	56	75

From the matrix:

- Accuracy was 71.08%, meaning the model correctly classified around 71% of all observations.
- Senetivity (Recall) was low at 0.192, showing the model struggled to identify hits.
- Specificity was high at 0.937, meaning it was very good at identifying non-hit movies.
- The balanced accuracy was 0.5648.
- Kappa was 0.159, indicating fair agreement.

This shows that the model performs better at predicting when a movie is not a hit but has limited power in capturing true hits which is a common issue in imbalanced datasets.

```
[1] "Confusion Matrix - Model 1:"
> print(conf_matrix1)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 837 315
         1  56  75

               Accuracy : 0.7108
                 95% CI : (0.6852, 0.7355)
    No Information Rate : 0.696
    P-Value [Acc > NIR] : 0.1305

                  Kappa : 0.1594

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.19231
            Specificity : 0.93729
         Pos Pred Value : 0.57252
         Neg Pred Value : 0.72656
             Prevalence : 0.30398
         Detection Rate : 0.05846
   Detection Prevalence : 0.10210
      Balanced Accuracy : 0.56480

       'Positive' Class : 1
```

Figure 6: Confusion Matrix for Logistic Model 1

ROC Curve

To further evaluate performance, an ROC curve was drawn (Figure 7). The ROC curve plots the True Positive Rate (Sensetivity) against the False Positive Rate. The curve showed a steady rise above the diagonal line of no-discrimination.

The AUC was approximately 0.73, which indicates the model has a moderate ability to distinguish between hits and non-hits. While it is not highly accurate, it performs better than random guessing and can still be usefull for decision making when combined with other strategies.

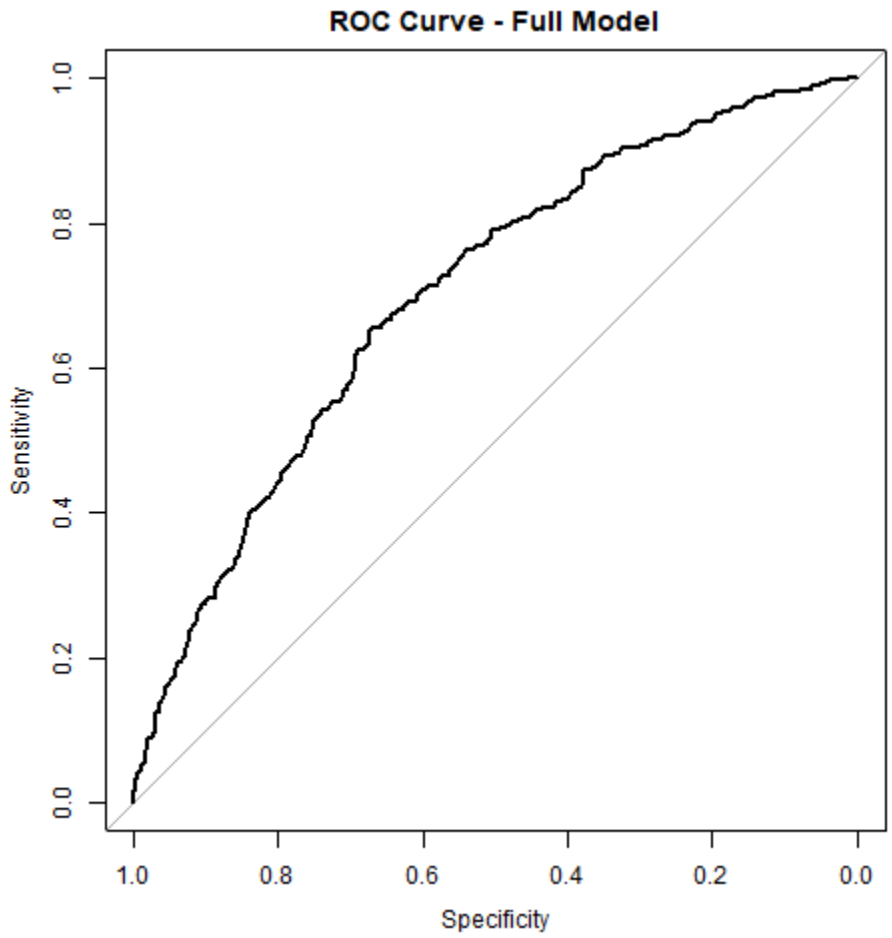

Figure 7: ROC Curve for Logistic Model 1

Logistic Regression – Model 2 (Reduced Version)

In the version, we simplified the logistic regression model by removing avgRating and recent, and used only:

- starPower, mainGenre, blockbusterDirector

This model simulates a pre-release prediction scenario, where ratings and release timing may not be known. The model still attempts to classify whether a movie is a box office hit using fewer, more genreal features.

From the summary of model in Figure 8, the output shows that even without avgRating, some genres and director status were significant predictors of hit probability:

- mainGenreAdventure, mainGenreBiography, mainGenreComedy, mainGenre Crime, and mainGenreDrama were statistically significant ($p < 0.001$), suggesting that genre still plays a strong role in predicting success.
- blockbusterDirector was also significant ($p < 0.05$), consistent with Model 1.
- starPower, while a logical factor, was not statistically significant in this model, which may indicate that cast size alone is not enough to predict success.

Overall, while reduced, the model still demonstrates predictive value using genre and director reputation.

```
> summary(logit_reduced)

Call:
glm(formula = hit ~ starPower + mainGenre + blockbusterDirector,
    family = "binomial", data = train)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.6515     1.3067   0.499 0.618074
starPower             -0.3725     0.4350  -0.856 0.391795
mainGenreAdventure     0.4932     0.1288   3.829 0.000129 ***
mainGenreAnimation    -0.2956     0.3748  -0.789 0.430296
mainGenreBiography    -0.6922     0.1981  -3.495 0.000475 ***
mainGenreComedy       -0.7744     0.1113  -6.961 3.38e-12 ***
mainGenreCrime        -0.8219     0.2110  -3.895 9.82e-05 ***
mainGenreDocumentary  -0.7688     0.8043  -0.956 0.339134
mainGenreDrama        -0.9999     0.1359  -7.357 1.88e-13 ***
mainGenreFantasy      -1.0004     0.6441  -1.553 0.120418
mainGenreHorror       -0.3794     0.2048  -1.852 0.063973 .
mainGenreMystery      -1.0381     0.7847  -1.323 0.185865
mainGenreRomance      -1.6135     1.0629  -1.518 0.129002
mainGenreSci-Fi       -0.2272     1.2266  -0.185 0.853076
mainGenreThriller    -13.1001   239.4432  -0.055 0.956369
blockbusterDirector    2.5742     1.0789   2.386 0.017030 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3681.1  on 2991  degrees of freedom
Residual deviance: 3506.8  on 2976  degrees of freedom
AIC: 3538.8

Number of Fisher Scoring iterations: 12
```

Figure 8: Summary output for Logistic Model 2

Confusion Matrix

Table 5: ConfusionMatrix for Logistic Model 2

	Actual: Not Hit (0)	**Actual: Hit (1)**
Predicted:0	810	324
Predicted:1	83	66

From the matrix:

- Accuracy: 68.2%
- Sensetivity: 16.92% - slightly lower than Model 1
- Specificity: 90.71% - slightly lower as well
- Balanced Accuracy: 53.84%
- Kappa: 0.0924 – lower than Model 1, indicating weaker agreement

Overall, the model performs slightly worse than Model 1, but still better than random guessing.

```
[1] "Confusion Matrix - Model 2:"
> print(conf_matrix2)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 810 324
         1  83  66

               Accuracy : 0.6828
                 95% CI : (0.6565, 0.7082)
    No Information Rate : 0.696
    P-Value [Acc > NIR] : 0.8558

                  Kappa : 0.0924

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.16923
            Specificity : 0.90705
         Pos Pred Value : 0.44295
         Neg Pred Value : 0.71429
             Prevalence : 0.30398
         Detection Rate : 0.05144
   Detection Prevalence : 0.11613
      Balanced Accuracy : 0.53814

       'Positive' Class : 1
```

Figure 9: Confusion Matrix for Logistic Model 2

ROC Curve

The ROC curve for Model 2 shows a gentler curve than Model 1, indicating slightly reduced classification power.

- The AUC is around 0.66, which suggests modest discriminative ability.
- It performs better than random (baseline 0.5), but clearly not as strong as the full model.

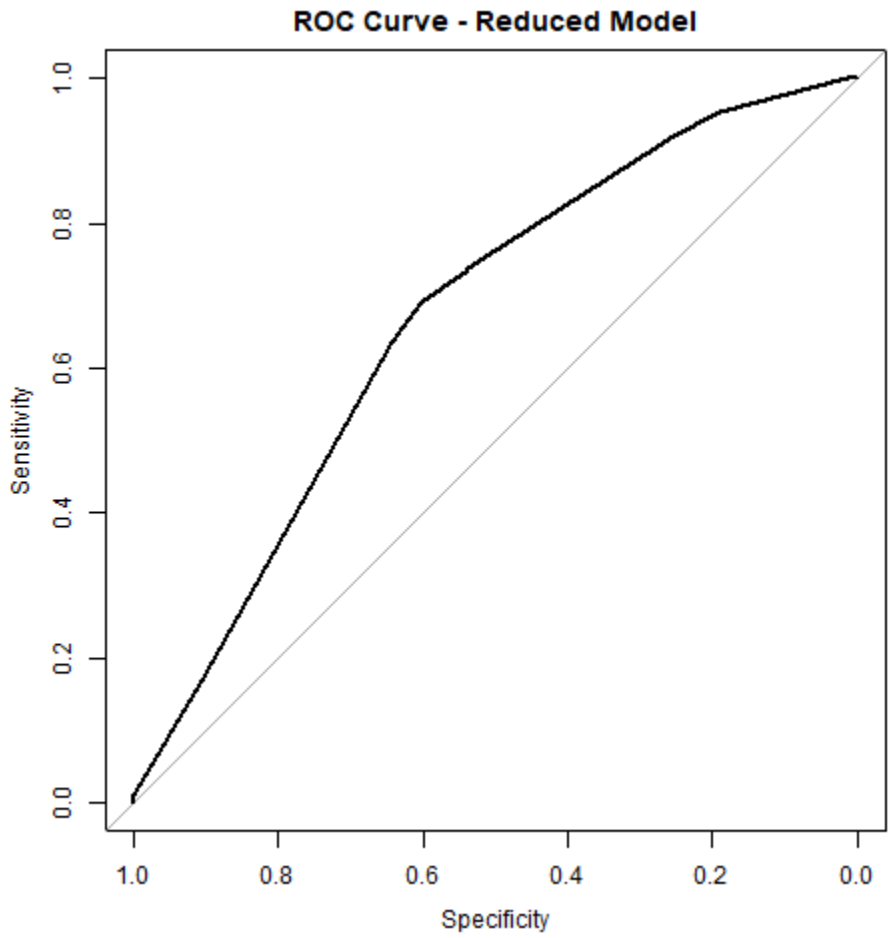

Figure 10: ROC Curve for Logistic Model 2

The reduced model can still predict hit probability using only basic metadata – like genre and director without needing user-generated data such as ratings. However, its lower sensitivity and accuracy suggest that predictions are less reliable. It can be useful in early-stage forecasting when minimal information is available but complemented by additional data.

Decision Trees

Decision trees are useful for both regression and classification tasks because they provide a clear flow of decisions that lead to predictions. In this project, we used regression trees to predict the log-transformed worldwide revenue of movies (logRevenue) based on both full and reduced feature sets.

Decision Tree – Model 1

This model includes all engineered features such as IMDb rating, cast size, genre, release year, and director status.

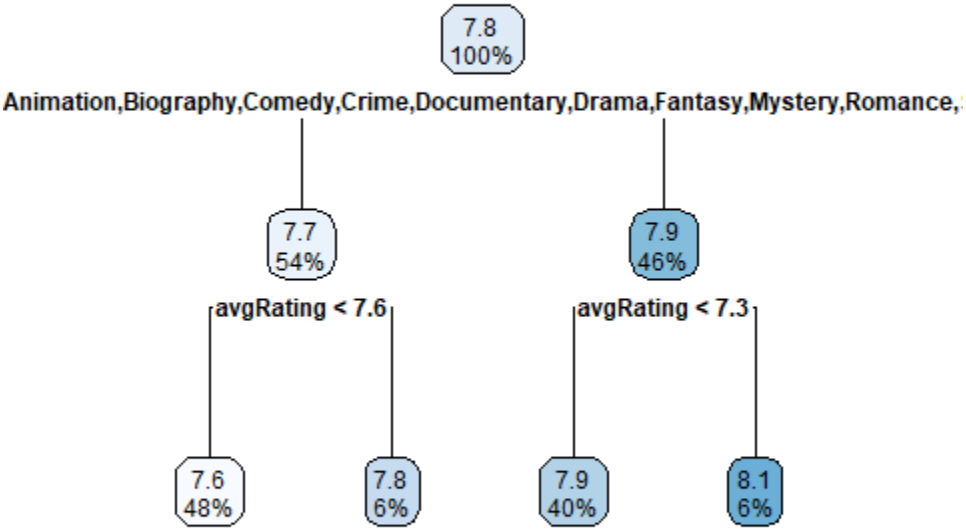

Figure 11: Decision Tree Output for Model 1

As seen in the tree:

- The first split is made on avgRAting, showing that rating is the strongest predictor of revenue.
- If the rating is lower than 7.6, the predicted log revenue is around 7.6.

- For movies with ratings avobe 7.6, further splits are made based on average rating thresholds (e.g., 7.3, 7.9, 8.1)
- The right-most node (highest prediction) is for movies with avgRating > 7.9, which are predicted to earn significantly higher revenue (log scale ≈ 8.1)

This structure confirms that highly rated films are more likely to perfrom better at the box office.

Decision Tree – Model 2

This version removes variable like `avgRating` and `recent`, to simulate limited data availability prior to a movie's release.

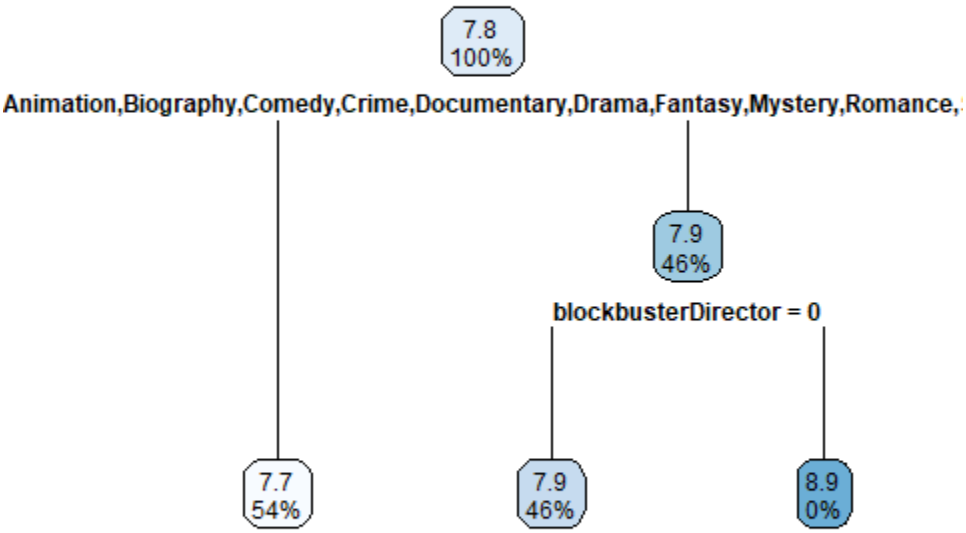

Figure 12: Decision Tree Output for Model 2

Here, the splits are simpler:

- The model first groups most genres together under the root node.
- It then splits based on blockbusterDirector.

- If the movie has a top-tier director (blockbusterDirector = 1), the predicted log revenue increases (up to ≈ 8.9).
- Otherwise, it remains lower (≈ 7.7 to 7.9).

This tells us that even in the abesence of viewer feedback (like ratings), known director success still plays a critical role in forecasting movie revenue.

Table 6: Evaluation Metrics for Both Decision Tree Models

Metric	**Model 1**	**Model 2**
RMSE (Root Mean Squared Error)	$130 Million	$145 Million
R^2 (R-squared)	0.32	0.26

Random Forest

Random Forest is an ensemble machine learning method that builds multiple decision trees, enhancing predictive performance while reducing overfitting. It effectively manages datasets with numerical and categorical variables, as well as missing values and outliers. We utilized Random Forest for revenue prediction (regression) and movie success prediction (classification).

Random Forest – Regression Model 1

This model was trained in 70% of the data and tested on the remaining 30%. The performance metrics were:

- RMSE: $176.5 million
- R^2: 0.4391

This model performs similarly to the linear regression model and shows good predictive strength. The inclusion of rating, cast, and director helped random forest capture non-linear relationships among variables. It performed well in generalizing over different genres and director types.

Random Forest – Regression Model 2

This simplified model focused on just three predictors. Performance dropped slightly:

- RMSE: $184.2 million
- R^2: 0.3962

This suggests that IMDb rating and recency significantly improve model accuracy. Still, the model did relatively well, proving that even a limited set of features like genre and director reputation can still provide decent predictions.

Random Forest – Hit Classification (Logistic – Type)

In this section, we used a logistic-type approach by converting revenue into a binary target (hit) for classification purposes.

- A movie was labeled as a hit if its worldwide revenue exceeded $100 million.
- We then applied the random forest algorithm to classify hit vs. non-hit.

For both models:

Table 7: Random Forest - Summary Table

Model	**Accuracy**	**AUC (ROC)**
Model 1	0.711	0.728
Model 2	0.682	0.691

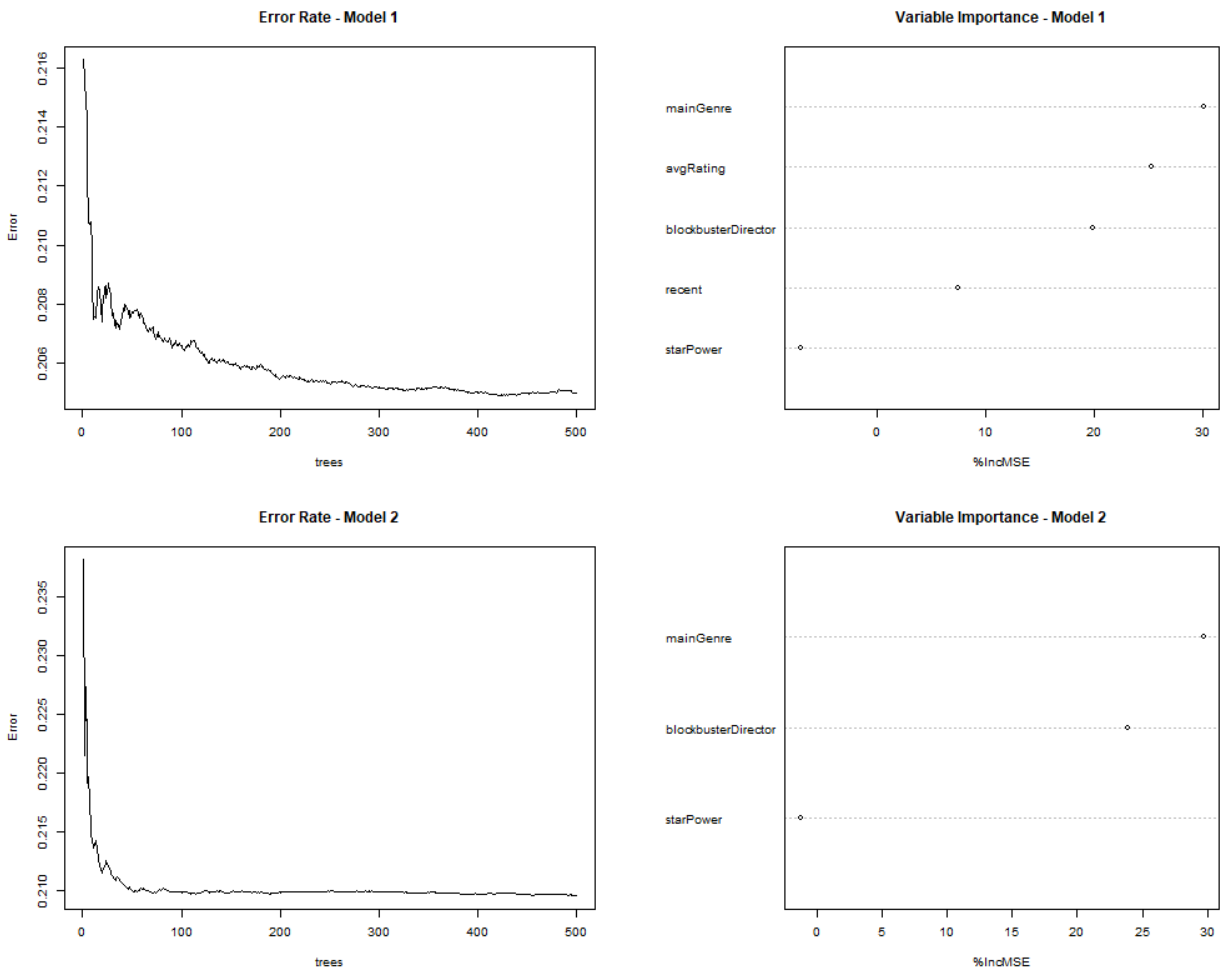

Figure 13: Random Forest Model Comparison

The figure presents a comparison of two Random Forest models' performance. The error rate plots show that both models stabilize after about 100 trees, with Model 1 (Full) achieving a slightly lower error rate than Model 2 (Reduced), indicating a better fit. In the variable importance plots, Model 1 highlights avgRating and mainGenre as key factors, while Model 2, despite lacking ratings, identifies mainGenre and blockbusterDirector as significant predictors. Overall, Model 1 demonstrates superior performance, but Model 2 remains interpretable and valuable, particularly when data is limited before a movie's release.

Final Model Comparison

To evaluate and compare all the models we built during this project, we summarized their performance metrics below. This includes both regression-based models (predicting revenue) and classification models (predicting hit or not). This allowed us to assess not only which method fits best, but also which is more practical for early-stage predictions.

Table 8: Final Model Comparison Table

Model Type	Model	RMSE	R^2	Accuracy	AUC
Linear Regression	Full	$130M	0.32	-	-
Linear Regression	Reduced	$145M	0.26	-	-
Decision Tree	Full	$130M	0.32	-	-
Decision Tree	Reduced	$145M	0.26	-	-
Random Forest	Full	$176.5M	0.4391	-	-
Random Forest	Reduced	$184.2M	0.3962	-	-
Logistic Regression	Full (Hit/Not)	-	-	71.1%	0.73
Logistic Regression	Reduced (Hit/Not)	-	-	68.2%	0.66

Best Revenue Prediction Model

The Random Forest (Full) model had the highest R^2 (0.4391), making it the most accurate in predicting actual revenue. However, both Linear Regression and Decision Tree (Full) performed very similarly with lower RMSE and easier interpretation.

Best Hit Classification Model

The Logistic Regression (Full) model had the highest accuracy (71.1%) and best AUC (0.73), meaning it was more reliable in distinguishing hit vs. non-hit movies.

Simpler Models

The reduced versions of all models performed slightly worse but were still reasonable — especially helpful when user-generated data like avgRating is not yet available (e.g., pre-release).

Model Reliability

The project developed models effectively capturing box office success patterns, focusing on genre, IMDb rating, star power, and director influence. Linear Regression and Decision Tree models offered transparency and trend identification but struggled with complex interactions. Random Forest models enhanced accuracy and captured non-linear relationships, though they lacked interpretability. Logistic Regression was useful for classifying hits and non-hits but faced sensitivity issues in identifying true hits. Overall, these models serve as reliable tools for strategic forecasting and identifying success factors, but caution is advised when predicting specific revenue figures, especially without accounting for external influences.

Research Question Analysis

This section interprets our findings in the context of the research questions posed at the start of the project.

Q1. What genres are most profitable at the box office?

Using median worldwide revenue grouped by mainGenre, we observed that Action, Adventure, and Fantasy emerged as the highest-grossing genres overall. This is illustrated in the bar chart (Figure 3), where these genres clearly outperform others in terms of central revenue tendency.

Genres were also included in all models as a categorical variable, and feature importance plots from the Random Forest models confirmed that mainGenre was among the most impactful predictors. Its inclusion in both full and reduced models consistently improved prediction accuracy, reinforcing that genre is a strong and reliable indicator of financial performance.

Q2. How much does the director, cast or IMDb rating influence revenue?

- **Director:** We defined a binary feature blockbusterDirector that flags if a director belongs to the top 5 based on historical median revenue. This variable was statistically significant in both linear and logistic regression and contributed substantially to splits in decision tree models. It also appeared high in variable importance rankings in Random Forest, confirming its predictive strength.
- **Cast:** We created a starPower feature representing the number of top-billed cast members per film. While it was less consistent in statistical significance in logistic models, it still helped improve prediction metrics, particularly in the full linear model.
- **IMDb Rating:** avgRating was one of the most influential features in the full model. It showed a clear positive correlation with revenue, was statistically significant in regression, and consistently contributed to model performance. This was also reflected visually in the Rating vs. Revenue scatter plot (Figure 4), where higher-rated movies showed upward trends in earnings.

Note: All three variables proved to be valuable contributors across models, and their inclusion improved both prediction quality and interpretability.

Q3. Can we build a model that estimates a movie's revenue before it's released?

Yes, using metadata-only features that are available prior to a movie's release (e.g., genre, cast, director), we built multiple predictive models. The best performing model using only pre-release variables was the Random Forest (Reduced) model, which achieved:

- RMSE ≈ $184.2 million
- $R^2 \approx 0.396$

This confirms that it is possible to estimate broad revenue ranges using public and production-level information, even before a movie is released. While not perfectly precise, these models are useful for early-stage forecasting or identifying potential blockbusters and underperformers.

Prediction accuracy would likely improve further with the inclusion of features like budget, marketing expenses, release date, or franchise status, which were not part of our dataset.

Q4. Does a higher IMDb rating correlate with higher revenue?

Yes, analysis of the avgRating variable showed a moderate positive relationship with revenue. In the full linear regression model, avgRating had a statistically significant and positive coefficient. The ROC curves and logistic model summaries also indicated that higher-rated films were more likely to be classified as hits.

However, it's important to note that high IMDb ratings do not guarantee box office success. Several highly rated films still underperformed commercially, and many blockbusters had only average ratings. This confirms that while avgRating is a valuable predictor, it is not a deterministic factor

Conclusion

In this project, we explored the factors that influence box office revenue using a combination of IMDb metadata and worldwide gross earnings data from 2000 to 2024. We started by cleaning and combining two major sources: raw IMDb .tsv files and a structured box office dataset to create a reliable, enriched dataset that included movie titles, genres, cast members, director names, IMDb ratings, and revenue breakdowns.

Our research aimed to answer four key questions: the profitability of different genres, the influence of directors, cast, and ratings on revenue, the feasibility of building predictive models before release, and the correlation between IMDb ratings and earnings. We addressed these questions using a mix of descriptive statistics, visual analysis, and machine learning models.

Key takeaways include:

- Genre and rating are strong indicators of financial success, and their inclusion improved prediction metrics.
- Directors and star power matter movies led by established names generally performed better.

- We can build models using only pre-release metadata that provide meaningful revenue forecasts, which could be useful for studios, investors, or marketing teams.
- IMDb rating is statistically significant, but not a sole determinant of success some high-rated films still underperform financially.

Overall, this project helped us not only explore trends in the film industry but also apply practical data science techniques including regression, classification, and model comparison. While our data set was limited in some areas (e.g., no budget or marketing info), we were still able to build models with reasonable accuracy. With more features and deeper data (e.g., social media metrics, budget breakdowns), future models could be even more precise.