Analyzing Movie Success: The Impact of Credits, Revenue, and Audience Perception

https://github.com/Dennis-Ding1/movie-success-analysis/tree/main/Midterm

Haochen Ding

1. Introduction

Understanding what drives a movie's success is a crucial question in the film industry. Factors such as genre, budget, cast, and audience perception play a significant role in determining a movie's box office performance. This study using data collected from The Movie Database (TMDB) API and Kaggle. Aims to explore the relationship between movie credits, box office revenue, and audience ratings and perception, incorporating both numerical and text-based analysis. Specifically:

How do movie credits (e.g., genre, cast, production) influence box office revenue and audience perception?

To address this, I propose the following hypotheses:

- Credits and Revenue: Larger casts and larger production scale contribute to higher box office earnings. Higher budgets generally lead to greater box office revenue, but the rate of increase diminishes as budgets grow, diminishing returns.
- Audience Reviews and Perception: Sentiment analysis of reviews might reveal whether higher audience ratings align with revenue, or if some films succeed despite mixed audience perception.

By integrating credit-based attributes, financial performance, and text analysis, this study aims to identify key factors driving a movie's success in both revenue and audience reception.

2. Method

2.1 Data collection & Merging

For this analysis, I used four datasets: three sourced from TMDB API and one from Kaggle. Since TMDB does not provide box office revenue data, and unlimited access to the IMDb API is not available, I opted to use an existing Kaggle dataset that contains revenue, budget, IMDb ratings, and additional metadata (https://www.kaggle.com/datasets/alanvourch/tmdb-movies-daily-updates/data).

From TMDB api (https://api.themoviedb.org/), I collected:

1. The top 2000 highest-TMDB-rated movies, with related movie credits.
2. Genre IDs to genre names look up table.
3. Audience reviews (up to 10 per movie).

Datasets were merged with kaggle dataset, using movie IDs from TMDB as link to perform left join (TMDB dataset as base), ensuring proper integration of financial, genre, and audience perception data. The final dataset includes 2000 conversation, and 26 variables:

- Movie details: title, release date, original language, cast, genre etc.
- Box office performance and budget.
- Audience ratings (TMDB and IMDB).

Audience reviews are in a separate dataset, this is for faster performance. It contains 5026 observations (reviews).

2.2 Data cleaning

To ensure data usability, I performed several preprocessing steps on the dataset. This included extracting relevant numerical features, handling missing values, and filtering unrealistic financial data.

1. Feature Engineering

Some columns in the dataset contained comma-separated lists (e.g., production companies, cast, genres, and production countries). To make these attributes more analytically useful, I transformed them into numerical features representing the number of elements in each category:

- num_production_companies: The number of production companies involved in each movie.
- num_cast: The number of credited cast members.
- num_genre: The number of genres assigned to a movie.
- num_production_countries: The number of countries associated with production.
- first_genre: The most significant genre of movies. This is for data model analysis, but will not be used in text analysis. Then I use genre id and name look up dataset to cover id to actual name of the genre.

2. Handling Missing and Unrealistic Values

To ensure meaningful analysis, I filtered out movies with missing or unrealistic financial data:

- Movies with NA values in revenue or budget were removed, as these are essential for analyzing box office performance.
- Movies with revenue or budget less than 100 were excluded, as such values are likely incomplete or inaccurate records.

2.3 Analyze methodology

To explore the relationship between movie credits, box office revenue, and audience perception, I employ a combination of numerical modeling and text analysis. The study is structured into two key analytical components:

1. Analyzing the Influence of Movie Credits on Box Office Revenue

- Linear Regression: A multiple linear regression model is used to estimate the individual contributions of key factors, such as budget, cast size, production scale, and genre, to box office revenue. This model provides an interpretable framework to assess whether larger productions and higher budgets are associated with increased revenue. Additionally, the regression model helps evaluate the presence of diminishing returns.
- Random Forest: As revenue-generation mechanisms may not always follow a linear pattern, a random forest model is applied to account for non-linear dependencies and higher-order interactions. This method enables the identification of budget thresholds where revenue growth plateaus and potential interactions that may not be captured by a linear approach. By leveraging variable importance measures, random forests further aid in determining which credit-related features have the greatest influence on box office performance.

2. Examining Audience Perception Through Text Analysis

- Word Frequency & TF-IDF (Term Frequency-Inverse Document Frequency) Analysis: Identifies the most commonly used words in audience reviews across different genres and revenue categories while also using TF-IDF to highlight words that uniquely define high-grossing vs. low-grossing films. This combined approach helps uncover both general themes and distinctive language patterns.
- Clustering of Reviews: Uses k-means clustering on reviews to identify latent audience discussion patterns. Clusters may reveal themes such as critical praise vs. commercial appeal, showing how audience perception differs for successful vs. unsuccessful films.

3. Preliminary Results

3.1 Data Summary

Table 1: Summary of Numerical Variables

Variable	Min	Max	Mean	Variance	NA_Count
tmdb_rating	7.3	8.708000e+00	7.675043e+00	8.442960e-02	0
vote_count	300.0	3.715700e+04	5.426490e+03	3.790407e+07	0
revenue	881.0	2.923706e+09	1.601138e+08	8.322180e+16	0
budget	17300.0	4.600000e+08	3.492008e+07	2.864944e+15	0
imdb_rating	1.7	9.300000e+00	7.701266e+00	2.673886e-01	0
imdb_votes	341.0	3.017438e+06	3.238674e+05	1.580561e+11	0
num_production_companies	1.0	2.200000e+01	3.441350e+00	5.500147e+00	0
num_production_countries	1.0	9.000000e+00	1.500422e+00	8.076434e-01	0
num_cast	1.0	3.080000e+02	4.918734e+01	1.073960e+03	0
num_genre	1.0	7.000000e+00	2.575527e+00	9.995695e-01	0

Table 1 shows some statistical summary of the numerical variables used in this research. After data cleaning, they are all in reasonable range.

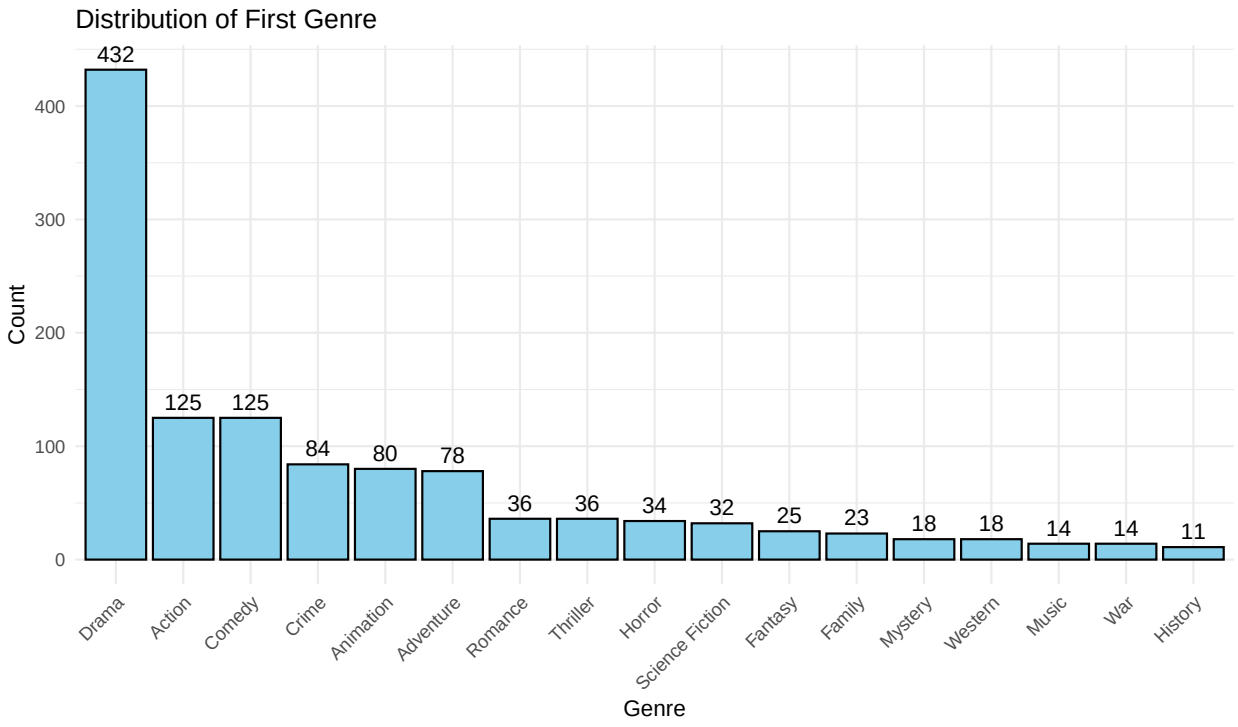

Figure 1: Distribution of First Genre

Figure 1 shows that Drama is the most common genre (432), followed by Action and Comedy (125 each). Less frequent genres include History (11), War (14), and Music (14), indicating a dominance of drama films in the dataset.

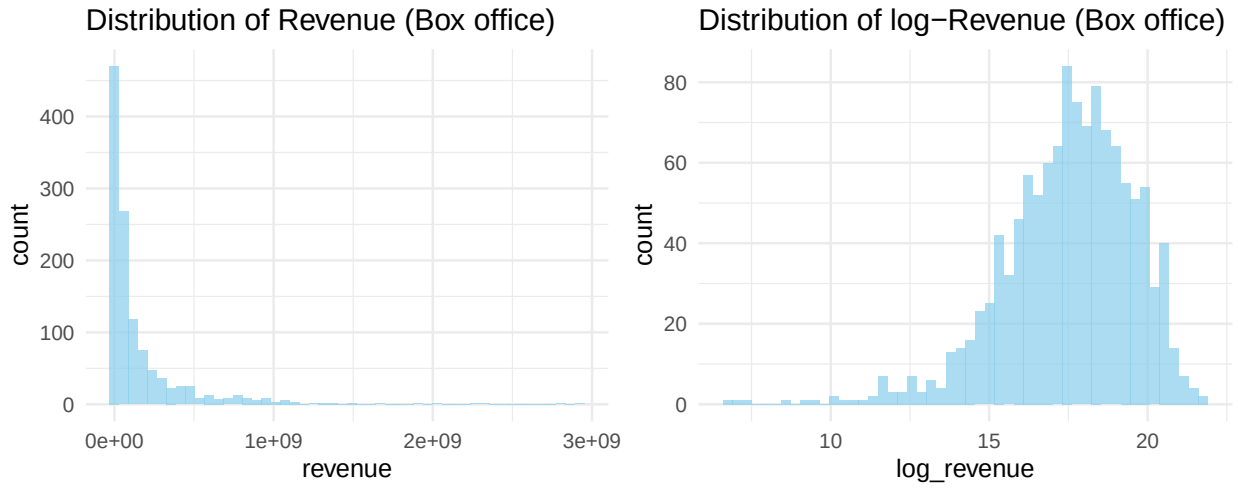

Figure 2: Distribution of raw and log-transformed revenue

Figure 2 shows the raw revenue distribution is highly skewed, with a long tail of extremely high values, making it difficult to analyze relationships effectively. Taking the log transformation of revenue reduces skewness, creating a more normal-like distribution. This transformation ensures regression model is meaningful.

3.2 Linear Regression

Table 2: Regression Results

term	estimate	std.error	statistic	p.value	significance
(Intercept)	10.8121445	1.2758640	8.4743705	0.0000000	***
budget	0.0000000	0.0000000	17.7789351	0.0000000	***
I(budget^2)	0.0000000	0.0000000	-10.5334350	0.0000000	***
num_production_companies	0.1225474	0.0264361	4.6356130	0.0000040	***
num_cast	0.0042119	0.0015939	2.6424292	0.0083422	**
num_production_countries	-0.2288664	0.0685493	-3.3387100	0.0008684	***
num_genre	-0.0088185	0.0563293	-0.1565533	0.8756242	
tmdb_rating	0.3988810	0.1976305	2.0183176	0.0437880	*
imdb_rating	0.3236167	0.1144066	2.8286532	0.0047549	**
first_genreAdventure	0.1035213	0.2310260	0.4480935	0.6541694	
first_genreAnimation	-0.2488709	0.2380323	-1.0455342	0.2959939	
first_genreComedy	-0.1631041	0.2142369	-0.7613258	0.4466172	
first_genreCrime	-0.3859256	0.2296569	-1.6804443	0.0931401	
first_genreDrama	-0.2176581	0.1770228	-1.2295489	0.2191154	
first_genreFamily	-0.3460928	0.3672450	-0.9424032	0.3461825	
first_genreFantasy	-0.1572979	0.3492649	-0.4503685	0.6525289	
first_genreHistory	-0.8333384	0.5043996	-1.6521392	0.0987768	
first_genreHorror	0.5590692	0.3178453	1.7589348	0.0788522	
first_genreMusic	0.3829636	0.4524470	0.8464276	0.3974888	
first_genreMystery	-0.6061662	0.4049667	-1.4968299	0.1347096	
first_genreRomance	0.1554092	0.3077424	0.5049978	0.6136563	
first_genreScience Fiction	0.0529635	0.3175792	0.1667727	0.8675780	
first_genreThriller	-0.6507685	0.3050760	-2.1331359	0.0331230	*
first_genreWar	-1.2448412	0.4510930	-2.7596110	0.0058781	**
first_genreWestern	-0.5117583	0.4130527	-1.2389662	0.2156087	

The multiple linear regression model examines the relationship between number of production companies / comtries / genre / cast, budget, and IMDB / TMDB ratings on log-transformed box office revenue. The model explains 42.5% of the variance, with an adjusted R-squared of 0.413, indicating a moderate fit.

Table 2 shows coeffiecent and their p values of the model:

1. Budget and Diminishing Returns:

- Budget has a strong positive effect on revenue, confirming that higher budgets tend to generate higher box office returns.
- Budget-squared is negative, indicating diminishing returns, meaning that after a certain point, increasing the budget leads to progressively smaller revenue gains.

2. Production and Cast Effects:

- Number of Production Companies has a positive impact, suggesting that films with multiple production companies tend to perform better.
- Number of Cast Members has a small but significant positive effect, meaning larger casts contribute slightly to higher revenue.

- Number of Production Countries is negatively correlated with revenue, implying that films with international co-productions may not always perform better financially.

3. Genre Influence:

The baseline genre is Action.

- Thriller and War have significantly lower revenues than Action films, suggesting they may not perform as well at the box office. However, their limited presence in the dataset suggests this trend may be influenced by sample size rather than an inherent genre effect.
- Other genres, such as Drama, Comedy, and Horror, do not show statistically significant differences from Action films.

4. Audience Ratings Impact:

- TMDB Rating and IMDB Rating both significant, and have a positive impact on revenue, indicating that higher audience ratings correlate with higher earnings.

3.3 Random Forest

A Random Forest model was trained to evaluate the factors influencing box office revenue based on budget, production details, audience ratings, and genre. The dataset was split into 80% training and 20% testing, and the model was trained with 100 trees.

Table 3: Random Forest Model Performance

Metric	Value
Mean Absolute Error (MAE)	9.128660e+07
Root Mean Squared Error (RMSE)	1.854698e+08
R-Squared (R^2)	6.180409e-01

Table 3 shows the Random Forest model explains 61.8% of revenue variance, with an MAE of $91.3M and an RMSE of $185.5M. While budget, genre, and ratings are key predictors, the high RMSE suggests missing factors like marketing or franchise status, indicating room for improvement.

Table 4: Random Forest Feature Importance

Feature	%IncMSE	IncNodePurity
budget	20.751154	3.425162e+19
first_genre	5.972379	7.742918e+18
imdb_rating	4.669885	5.811665e+18
num_production_countries	3.257114	1.184744e+18
num_production_companies	3.219107	2.919030e+18
tmdb_rating	2.987968	6.095374e+18
num_cast	1.816811	8.098030e+18
num_genre	1.303141	2.720649e+18

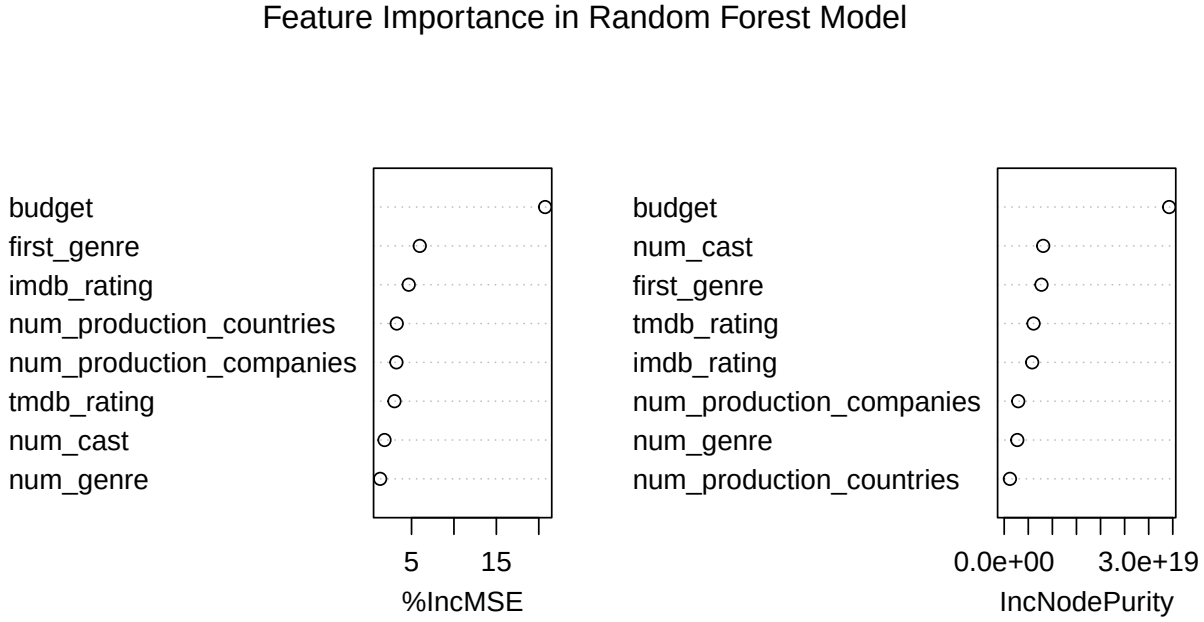

Figure 3: Feature Importance Plot

The Table 4 and Figure 3 shows the key factors affecting revenue:

- %IncMSE (Percentage Increase in Mean Squared Error) shows how much the model's prediction error increases when a given variable is randomly permuted. IncNodePurity (Increase in Node Purity) evaluates how much each feature reduces impurity (variance) in the decision trees.
- Budget is the most important predictor, with the highest %IncMSE (20.75) and Node Purity Contribution, confirming that higher budgets lead to higher revenue.
- First Genre (5.97% IncMSE) also plays a significant role, suggesting that genre type impacts revenue. However, this contrasts with the regression results, where most first_genre categories were statistically insignificant. This discrepancy suggests that genre does not have a strong independent effect on revenue but rather interacts with other factors like budget and cast size. Random Forest captures these nonlinear interactions, indicating that genre's influence is context-dependent, which may not be fully reflected in a linear model.
- IMDB rating (4.67%) and TMDB rating (2.99%) influence revenue, indicating that higher audience scores correlate with better financial performance.
- Production-related variables (e.g., number of production companies, number of cast members) contribute to revenue but are less impactful than budget and ratings.

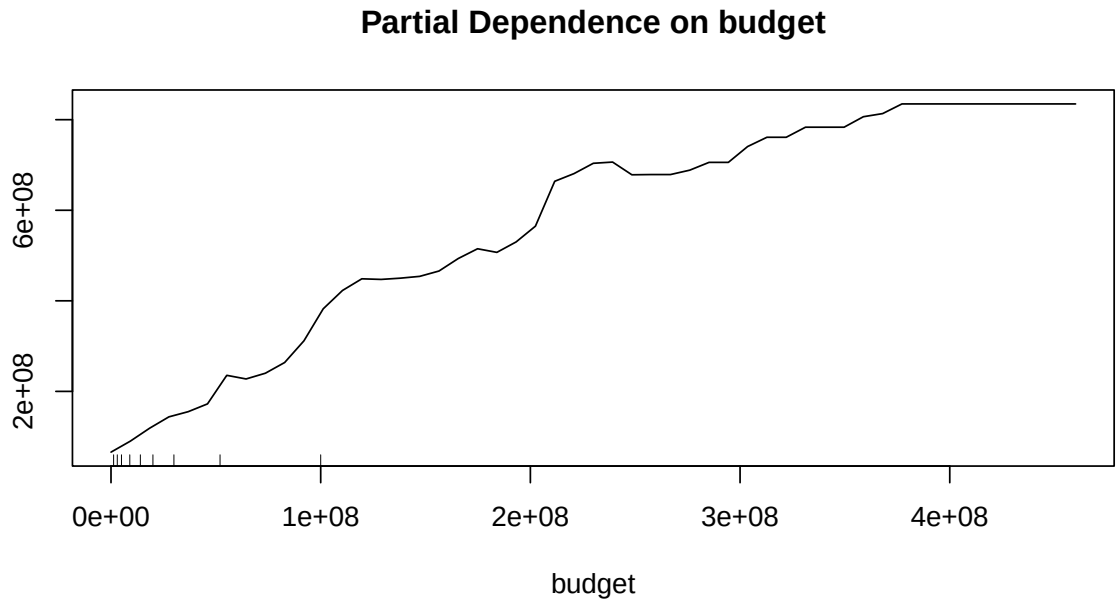

Figure 4: Partial Dependence on Budget

Figure 3 shows how predicted revenue changes with budget, holding other variables constant. The relationship is initially steep, indicating that increasing budgets significantly boost revenue. However, the curve flattens at higher budgets, suggesting diminishing returns—beyond a certain point, additional spending yields smaller increases in revenue. This aligns with our regression findings, where the quadratic budget term was negative, confirming that while budget is a key driver of revenue, its impact is nonlinear.

3.4 Text Frequency and TF-IDF anaylze

To be finish in Final Report.

3.5 Clustering for Review Analysis

To be finish in Final Report.

4. Summary

4.1 Finding so far

The analysis so far reveals several key insights into how movie credits and revenue interact:

- Budget is the strongest predictor of box office revenue, confirmed by both linear regression and Random Forest models. However, the impact is nonlinear, with diminishing returns at higher budgets, as seen in the quadratic term in regression and the flattening trend in the partial dependence plot of random forest.

- Audience ratings (IMDB and TMDB) are positively associated with revenue, suggesting that higher-rated movies tend to perform better financially. However, the impact is moderate, indicating that critical acclaim alone does not guarantee commercial success.
- The role of genre is complex. Regression results show that most genres are not individually significant, and the ones with significant is small portion of the data. While Random Forest assigns high importance to first_genre, suggesting that genre influences revenue through interactions with other factors like budget and production scale.
- Production and cast size contribute to revenue, but their effects are smaller than budget and ratings. The number of production companies has a positive effect, possibly indicating that higher production investment leads to greater distribution and marketing efforts.

4.2 Futher anaylze plan

The next steps will focus on text analysis of audience reviews. This will help uncover whether sentiment and key themes in reviews align with financial performance.

1. Text Frequency & TF-IDF Analysis

- Compute word frequency distributions across different genres to identify commonly used words in audience reviews.
- Apply TF-IDF (Term Frequency-Inverse Document Frequency) to highlight distinctive words that characterize high- vs. low-grossing films.

2. Clustering for Review Analysis

- Apply k-means clustering to group reviews to uncover latent themes in audience discussions.
- Identify whether certain review clusters correlate with box office success or audience ratings.

3. Sentiment Analysis & Correlation with Ratings and Revenue (If previous result shows interesting and worth exploring)

- Compute sentiment scores from reviews and correlate them with audience ratings and revenue.
- Test whether positive sentiment is a stronger predictor of revenue than numerical ratings.