

# Project Report

## The Cinematic Economy

**Analyzing the Impact of Budget, Star Prestige, Inflation, Genre, and Seasonality on Film Success**

**Author:** Can Sever

**Date:** 8th of January 2026

**Course:** DSA210

# 1. Executive Summary

This project investigates the multifaceted drivers of commercial success in the film industry. By enriching a primary dataset of 5,000 films with external economic indicators (CPI, Unemployment Rate), temporal data (US Federal Holidays), and prestige metrics (Oscar History), I have constructed a robust predictive framework.

The study employs statistical hypothesis testing (T-tests, ANOVA) and machine learning algorithms (Linear Regression, Ridge Regression, Random Forest). Key findings indicate that inflation-adjusted budget remains the dominant predictor of revenue. However, statistical tests confirm that seasonality (Holiday releases) and genre significantly impact earnings. The Random Forest model demonstrated superior performance over linear baselines, effectively capturing the non-linear relationships inherent in box office data.

## 2. Introduction

Predicting the financial success of a motion picture is a high-stakes challenge involving complex, interacting variables. Traditional models often rely solely on production attributes. This study expands the scope by integrating four distinct dimensions:

- **Economic Context:** How macro-economic health (Unemployment) and inflation influence purchasing power.
- **Star Prestige:** Quantifying the "star power" of cast and crew via historical Academy Award data.
- **Temporal Factors:** Analyzing the premium of releasing films during major holiday windows.
- **Categorical Attributes:** Evaluating the variance in revenue across different film genres.

The primary objective is to build a high-quality predictive model for **Real Box Office Revenue** while statistically validating specific hypotheses regarding these factors.

## 3. Data Acquisition and Engineering

### 3.1. Data Sources

The project utilizes a primary dataset enriched with four external sources to fulfill the "Data Enrichment" requirement:

- **Primary Movie Data (Kaggle/TMDB)**: Core attributes (Budget, Revenue, Cast, Release Date).
- **US CPI (Bureau of Labor Statistics)**: Used for calculating "Real" (inflation-adjusted) Budget and Revenue.
- **Unemployment Rate (Federal Reserve - FRED)**: Serves as a macro-economic context proxy.
- **Academy Awards (Kaggle)**: Used for calculating cumulative "Prestige Scores".
- **Holiday Calendar (US Gov Data)**: Used for flagging temporal release windows.

### 3.2. Preprocessing & Feature Engineering

- **Data Cleaning**: JSON columns (Genres, Cast) were parsed. Rows with negligible financial data (less than \$1,000) were removed to ensure data quality.
- **Inflation Adjustment (H1)**: Using CPI data joined by Release Year, nominal values were converted to 2016-adjusted dollars (Real\_Budget, Real\_Revenue), removing the distortion of time.
- **Star Prestige Index (H2)**: A custom algorithm calculated the cumulative number of Oscar wins for the Director and Top 3 actors *prior* to the film's release, preventing look-ahead bias.
- **Seasonality (H3)**: A binary feature named "Is\_Holiday" was engineered by cross-referencing release dates with a database of major US holidays (intervals of  $\pm 7$  days).

## 4. Exploratory Data Analysis & Hypothesis Testing

I formally tested four hypotheses to understand the statistical significance of the features.

### 4.1. H1: The Economic Effect

- **Hypothesis:** Higher budgets and lower unemployment rates correlate with higher revenue.
- **Analysis:** Correlation matrices revealed a strong positive correlation between Real Budget and Real Revenue. The relationship with Unemployment Rate was present but weaker, suggesting that while the economy matters, blockbuster appeal often defies economic downturns.

### 4.2. H2: Star Prestige Index

- **Hypothesis:** High "Star Prestige" (past Oscar wins) leads to higher revenue.
- **Analysis:** Scatter plots and correlation coefficients indicated a positive relationship. Films with high prestige scores often secure higher initial budgets, indirectly boosting revenue.

### 4.3. H3: The "Holiday" Effect

- **Hypothesis:** Films released during holidays earn significantly more.
- **Test:** Independent T-test.
- **Result:** The test rejected the null hypothesis. There is a statistically significant difference in mean revenue between holiday and non-holiday releases, validating the industry strategy of "tentpole" holiday releases.

### 4.4. H4: Genre Effect

- **Hypothesis:** Film genre creates a significant difference in average gross.
- **Test:** One-Way ANOVA (Analysis of Variance).
- **Result:** The ANOVA test yielded a p-value well below 0.05. Distinct revenue tiers were observed, with "Adventure" and "Action" genres significantly outperforming "Drama" and "Horror" in terms of average gross revenue.

## 5. Machine Learning Implementation

### 5.1. Model Setup

I have approached the revenue prediction as a regression problem.

- **Target:** Real Revenue
- **Features:** Real Budget, Unemployment Rate, Star Prestige Index, Is Holiday, Runtime, Vote Average.
- **Preprocessing:** Data was split (80% Training, 20% Testing) and standardized (StandardScaler) to ensure fair weight distribution for linear models.

### 5.2. Models Evaluated

1. **Linear Regression:** Used as a baseline to understand linear relationships.
2. **Ridge Regression:** Applied to mitigate multicollinearity between correlated features (e.g., Budget and Prestige).
3. **Random Forest Regressor:** An ensemble method used to capture non-linear patterns and interactions between features.

## 6. Results and Discussion

### 6.1. Model Performance

The evaluation metrics (R-Squared Score and RMSE) highlighted the superiority of ensemble methods for this dataset.

- **Linear / Ridge:** Provided a solid baseline but struggled with the massive variance in box office data (heteroscedasticity). Predictions tended to be conservative for blockbusters.
- **Random Forest: Best Performer.** Achieved the highest R-Squared score. The model successfully captured complex interactions, such as the compounding effect of a high budget combined with a holiday release.

### 6.2. Visual Analysis

- **Linearity vs. Complexity:** Visualizations showed that Linear Regression forced predictions into a straight line, failing to capture the exponential nature of "mega-hits."
- **Log-Scale Evaluation:** When analyzing Random Forest results on a Logarithmic Scale, the predictions formed a cohesive cluster around the diagonal, indicating that the model handles orders of magnitude (millions vs. billions) effectively, though some variance remains in the lower-revenue tail.

### 6.3. Feature Importance

Analysis of the Random Forest feature importance confirmed the EDA findings:

1. **Real Budget:** The single most predictive feature. Money begets money in the film industry.
2. **Vote Count / Popularity:** Served as strong proxies for marketing reach and audience engagement.
3. **Runtime & Prestige:** Contributed meaningfully but were secondary to financial inputs.

## 7. Conclusion

This project successfully demonstrates that film success is not random but strongly correlated with identifiable factors. By integrating economic context (Inflation) and temporal markers (Holidays), I improved the explainability of the model.

While Budget remains the king of predictors, the statistical tests proved that Strategic Release Timing (Holidays) and Genre Selection are statistically significant levers for maximizing revenue. The transition from Linear Regression to Random Forest highlighted the non-linear nature of the cinematic economy, where the "blockbuster effect" creates exponential rather than linear returns.

## 8. Limitations and Future Work:

- **Limitations:**
  - **Linearity Issues:** Linear Regression models struggled with the massive variance (heteroscedasticity) in box office data, often underpredicting "mega-hits" by forcing predictions into a straight line.
  - **Geographic Bias:** The economic indicators (CPI, Unemployment) and holiday calendars are US-centric, which limits the model's ability to fully explain global box office performance.
- **Future Work:**
  - **Global Expansion:** Future iterations could incorporate global economic data to better predict international revenue.
  - **Sentiment Analysis:** Integrating Natural Language Processing (NLP) on movie plots or early social media sentiment could provide a new dimension of "hype" quantification that is currently missing from the tabular data.
  - **Advanced Modeling:** Exploring deep learning techniques to better handle the "long tail" of lower-revenue films where variance remains high.