Projecting Blockbusters: A Machine Learning Framework for Movie Greenlighting at NeoStudio Pictures

Saad Iftikhar[1], Talal Naveed[1], and Ahmed Arkam Mohamed Faisaar[1]

[1]Predictive Analytics Division, NeoStudio Pictures

December 2024

Abstract

The "Greenlight" decision in the film industry is characterized by high capital intensity and extreme uncertainty. At NeoStudio Pictures, we address this by developing a dual-model machine learning framework that utilizes pre-release metadata to forecast commercial outcomes. Using a cleaned subset of the TMDB 5000 dataset ($N = 3,229$), we implemented a LightGBM classification system to identify "Hits" ($\geq$\$100M revenue) and an XGBoost regression system for revenue magnitude forecasting. Our classification model achieved an ROC-AUC of 0.9762 and a recall of 95.83%, while the regression model attained an R^2 of 0.8577. This paper details the feature engineering pipeline, temporal validation strategy, and business applications of these models in risk management. Importantly, we position this system as decision support rather than an automated gatekeeper, emphasizing probabilistic reasoning over deterministic approval.

Contents

1 Introduction

1.1 Business Motivation

The film industry operates on a "hit-driven" economy. For a mid-size studio like **NeoStudio Pictures**, a single failed blockbuster can jeopardize the company's annual solvency. Conversely, identifying a sleeper hit early allows for optimized marketing spend and distribution leverage.

Beyond direct revenue, greenlighting decisions influence opportunity cost: every approved project displaces alternative scripts, talent packages, and release windows. Poor decisions therefore compound risk not only through losses, but through missed upside.

The primary challenge is *information asymmetry*: decisions must be made years before release based on scripts and talent attachments. This project seeks to bridge the gap between creative intuition and statistical probability by providing a repeatable, data-driven framework for greenlighting decisions. Rather than replacing executive judgment, the framework structures it.

1.2 Problem Definition

We define the predictive task in two distinct tiers:

1. **Binary Classification**: Predicting if a project will exceed the $100M worldwide revenue threshold.
2. **Continuous Regression**: Forecasting the exact log-transformed revenue to estimate ROI.

The classification task supports portfolio-level decisions (e.g., which projects to prioritize), while the regression task informs budgeting, marketing scale, and distribution strategy.

2 Data Acquisition and Preparation

2.1 Data Sources

The study utilizes the TMDB 5000 dataset, merged across movie attributes and credit metadata.

- **Features**: Budget, Genre, Production Companies, Release Dates.
- **Credits**: Full cast lists and director identities.

TMDB is particularly suitable for this task because it provides standardized global revenue figures and structured metadata across decades of releases.

2.2 Rigorous Data Cleaning

To ensure the model reflects reality, we applied strict filters:

- **Financial Filtering**: Removed movies with \$0 budget or revenue (common in missing data).
- **Temporal Consistency**: Standardized release dates into datetime objects to extract seasonal features.
- **Outlier Handling**: Extreme but implausible budget-to-revenue ratios were reviewed to avoid data-entry artifacts.
- **Final Dataset**: A high-fidelity sample of 3,229 movies ranging from 1916 to 2016.

This aggressive cleaning step trades dataset size for realism, aligning the model with actual executive decision contexts.

2.3 Class Distribution

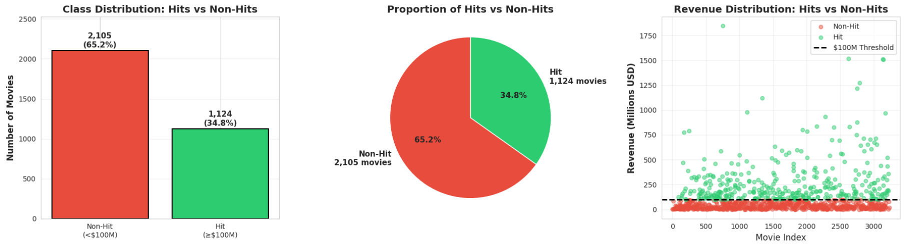

Figure 1: Distribution of hit versus non-hit films in the cleaned dataset.

3 Methodology

3.1 Feature Engineering: The "Greenlight" Levers

We transformed raw metadata into decision-relevant features:

1. **Talent Track Record**: For every director, we calculated the *Director Hit Rate* (percentage of previous films $\geq$\$100M) and *Mean Log Revenue*. This creates a quantitative proxy for a director's "bankability," capturing both consistency and upside.
2. **Temporal Features**: Binary flags for *is_summer* and *is_holiday* encode market congestion, school calendars, and historical audience behavior.
3. **Financial Scaling**: We applied log-transformation to budget: $f(x) = \log(1 + x)$ to stabilize variance across "micro-budget" and "mega-blockbuster" tiers. This ensures the model learns relative scale rather than raw dollar magnitude.

4. **Institutional Proxy**: A *has_big_studio* flag was created by identifying the top 10 historical distributors. This captures downstream advantages in marketing reach, international distribution, and theater access.

3.2 Budget–Revenue Relationship

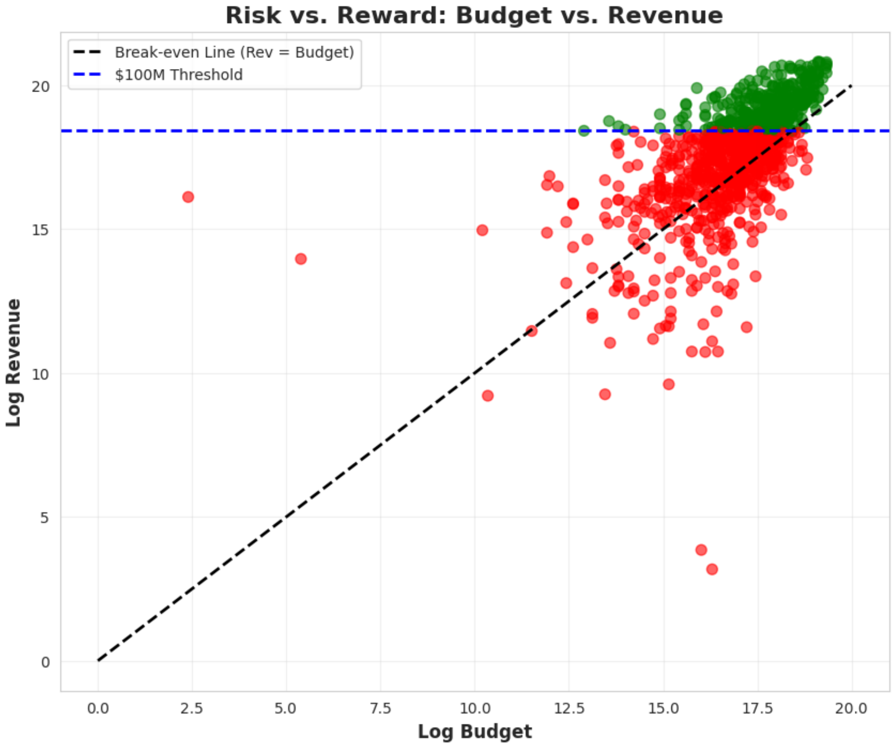

Figure 2: Relationship between production budget and worldwide revenue.

3.3 Temporal Validation Strategy

Unlike standard cross-validation, movie data is time-dependent. We used a **Temporal Train-Test Split** to simulate a real executive scenario:

- **Training**: Movies released before 2010 ($N \approx 2,200$).
- **Validation**: 2010–2012.
- **Final Evaluation (Test)**: 2013–2016.

This approach prevents "training on the future" and produces conservative, deployment-ready performance estimates.

4 Analysis and Results

4.1 Classification Results (Hit Prediction)

The LightGBM model demonstrated exceptional separation between hits and non-hits, even under class imbalance conditions.

Metric	LightGBM Performance
ROC-AUC	0.9762
Accuracy	90.21%
Recall (Hit Class)	95.83%
F1-Score	0.8980

Table 1: Model performance on the post-2013 Test Set.

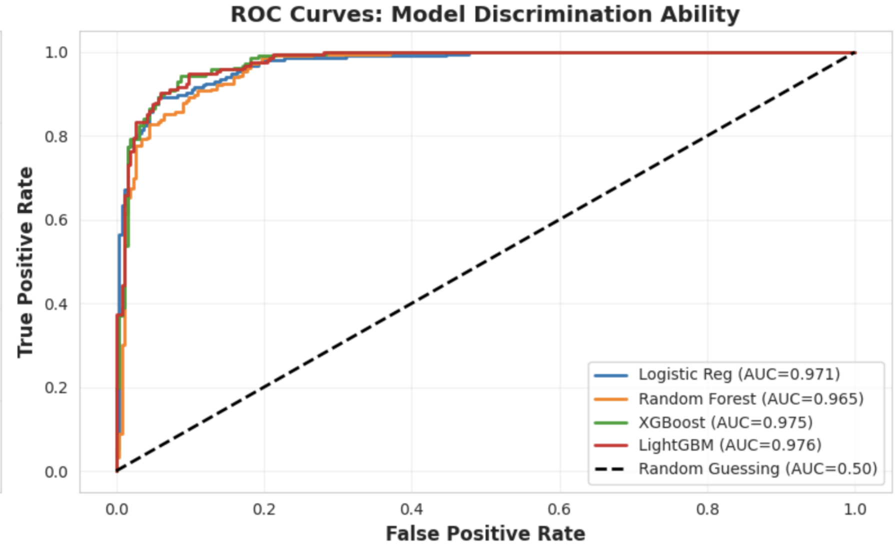

Figure 3: ROC curve for the LightGBM hit classification model.

High recall is particularly valuable in this context: missing a true hit represents a lost blockbuster opportunity, whereas greenlighting a false positive can often be mitigated through budget controls.

4.2 Regression Results (Revenue Forecasting)

For predicting the magnitude of revenue, the XGBoost Regressor achieved an R^2 of 0.8577. This indicates that over 85% of revenue variance is explainable using pre-release signals alone.

While individual film outcomes remain noisy, the regression model provides strong directional guidance for financial planning.

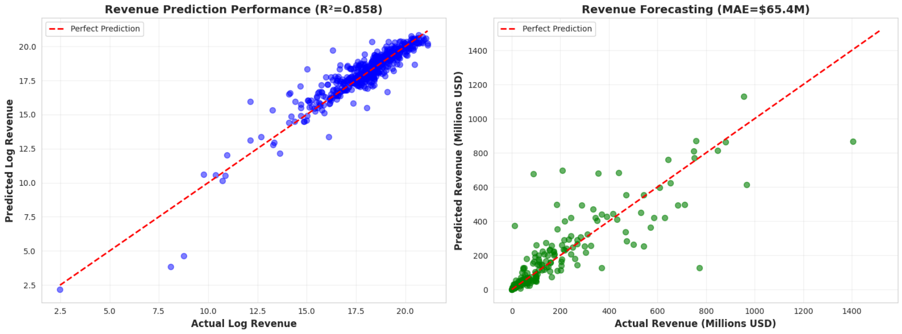

Figure 4: Predicted versus actual log-transformed revenue on the test set.

5 Discussion: Executive Insights

- **Finding A — Budget Sensitivity**: Higher budgets increase hit probability but also amplify downside risk. Variance grows faster than mean return, reinforcing the need for disciplined scaling.
- **Finding B — The Director Effect**: Director track record dominates other creative signals. This supports the industry intuition that proven leadership reduces execution risk.
- **Finding C — Strategic Windows**: Summer releases exhibited a 1.9x higher hit likelihood. Timing acts as a force multiplier rather than a cosmetic scheduling decision.
- **Finding D — Structural Advantage**: Big-studio backing materially shifts baseline probabilities, highlighting the importance of distribution and marketing infrastructure.

6 Business Application: Streamlit Implementation

To operationalize these insights, we developed an interactive Streamlit dashboard enabling scenario-based exploration. Executives can:

1. Input proposed budget and genre.
2. Select a director and immediately observe probability shifts.
3. Adjust release timing to assess strategic tradeoffs.

This transforms the model from an analytical artifact into a practical decision instrument.

7 Conclusion and Future Work

Our framework provides a statistically grounded baseline for greenlighting decisions under uncertainty. By translating historical patterns into actionable probabilities, NeoStudio can systematically reduce uninformed risk without suppressing creative ambition.

Future extensions include:

- **NLP-Based Script Analysis**: Incorporating screenplay summaries using Transformer models.
- **Inflation Adjustment**: Normalizing historical revenue to constant dollars.
- **Streaming Metrics**: Integrating SVOD performance to capture modern success pathways.