DERIVING PREDICTIVE MODELS FOR FILM SUCCESS USING STATISTICAL AND MACHINE-LEARNING METHODS

THE COLLEGE OF WOOSTER

Patricia Chen, Advised by Dr. Robert Kelvey

ABSTRACT

Predicting the success of films has been a challenge for the movie industry, where billions of dollars are invested into projects whose outcomes remain highly uncertain. This independent study explores whether pre-release factors are helpful in forecasting film success through mathematically derived statistical and machine learning techniques. Utilizing data from IMDb, Rotten Tomatoes, Metacritic, the Academy awards, the Golden Globes, and budget and revenue records, a dataset was assembled to capture a range of film characteristics such as release month, runtime, MPAA rating, genre, and more. We developed two predictive models. The first one is a risk scoring model derived from logistic regression, where the regression's coefficients are transformed into an interpretable points-based score to predict breakeven proportion. The second approach uses the CatBoostRegressor, a gradient-boosting machine learning method designed to handle categorical variables to predict proportional revenue. Together, these two methods provide a greater understanding into how movie predictors shape their successes and portray the value of using both interpretable statistical frameworks and machine learning techniques.

PREDICTORS

Variable	Definition	Categories
Month	Release month	1–12
Runtime	Duration of the film	Short $<$ 90 min Medium $\leq$ 120 min Long $>$ 120 min
Rating	MPAA rating	G, PG, PG13, R
Distribution_Company	Top 10 distribution company	1 = yes, 0 = no
Budget	Film budget	3 quantile buckets: Low, Medium, High
Oscar_Director_History	Prior Oscars won by director	0, 1–2, 3+
Oscar_Actor_History	Prior Oscars won by actors	0, 1–2, 3+
GG_Director_History	Prior Golden Globes won by director	0, 1–2, 3+
GG_Actor_History	Prior Golden Globes won by actors	0, 1–2, 3+
Genres: Action...Western	Indicator for genre membership	1 = yes, 0 = no

Table 1: Variables & Definitions

RISK SCORING MODEL

A risk scoring model calculates a numerical value based off predictors (risks) that summarizes how "risky" or likely an outcome is. The foundation of the risk scoring model stems from logistic regression. We used this model to measure the probability of a film breaking even defined as 2.5 times the film's budget.

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

- Tested various logistic regression models using stepwise regression
- Observed each model's AIC & residual deviance

$$AIC = 2K - 2\ln(L) \qquad \text{Residual Deviance} = -2(\log L_{predicted} - \log L_{saturated})$$

- Using the best model, we multiplied the coefficients by 100, creating vector $\vec{v}$
- We created a matrix *M* representing the attributes per film *i*
- Multiplying $M\vec{v} = \vec{S}$, where $\vec{S}$ is a vector representing each film's risk score

CATBOOSTREGRESSOR MODEL

The CatBoostRegressor model is a machine learning model that uses gradient boosting to train itself to create predictions. We used this model to measure the proportional revenue that a film makes. This computation was completed in R Studio.

$$f_{loss}(F) = \frac{1}{N}\sum_{n=1}^{N}(y_n - F(x_n))^2$$

$$-\frac{\partial f_{loss}}{\partial F(x_n)} = \frac{2}{N}(y_n - F(x_n))$$

- Transformed the response variable logarithmically so its distribution is more normal
- Split data into a training and testing set, 80/20 respectively
- The model ran through 1000 sequential decision trees
- Predictions were generated based off the testing data on the same logarithmic scale
- Converted the response variable back into its original form

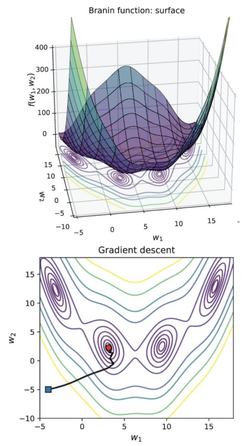

Figure 1: Visuals of gradient descent [1]

RESULTS

The risk scoring model has a positive correlation between the risk score we created and the proportion of movies that broke even.

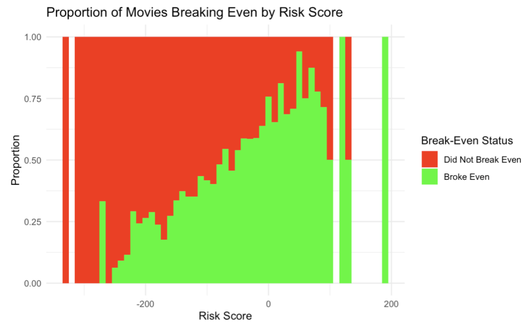

Figure 1: Proportion of Movies that Breakeven Relative to its Risk Score

We also divided the risk scores into five strata so we could further analyze the different attributes between the films with a higher versus lower risk score.

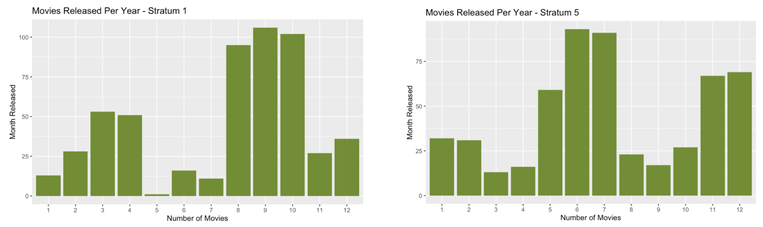

Figure 2: Differences between the distribution of Stratum 1 and Stratum 5's release month

Stratum 1

- Top 5 Genres: Drama, Crime, Action, Comedy, and Biography
- Runtime: one large spike with a median of 108
- Top 10 Distribution Company: 177 films
- Median Budget: $33.5 million

Stratum 5

- Top 5 Genres: Comedy, Adventure, Horror, Drama, and Thriller
- Runtime: a large spike at 90 minutes and a smaller spike at 130 minutes
- Top 10 Distribution Company: 305 films
- Median Budget: $15 million

Predictor	Feature Importance
Budget_cat	14.6413
Month	13.9222
Rating	10.8389
Runtime_cat	10.7091
GG_actors_cat	9.7637
Distribution_Company_cat	6.1837
Horror	4.1100
Mystery	3.5126
Comedy	3.3991
Oscars_directors_cat	2.9684

Table 2: CatBoost Feature Importance for Predicting Proportional Revenue (Top 10)

CatBoost shows solid predictive performance with errors well below the data's standard deviation, indicating meaningful fit despite the low R^2.

CatBoost also includes a function called Feature Importance that shows a significance value for how important an independent variable is in generating the predictions [2].

RMSE	MAE	SD	R Squared	R Squared (log)
6.52	2.56	11.17	0.07	0.11

Table 3: CatBoost Results

Both models were then evaluated on newer films that have already been released in Table 4.

Film	Release Year	Pred. Breakeven Prop.	Breakeven	Pred. Prop. Rev.	Prop. Rev.
Spiderman 3	2022	0.31	Yes	2.76	9.61
Top Gun	2022	0.50	Yes	2.85	8.76
Barbie	2023	0.30	Yes	3.21	9.65
Challengers	2024	0.00	No	1.06	1.74
Sinners	2025	0.56	Yes	2.09	3.70

Table 4: Model's Predictions vs Actual Outcome on Newer Movies

CONCLUSION

These two models illustrate how statistical and machine learning models can predict outcomes in the film industry, an industry that comes with lots of uncertainty. The risk scoring model offers transparency and interpretability while the CatBoostRegressor model provides an additional perspective by capturing nonlinear relationships. Using both of these models highlights the importance of incorporating multiple perspectives of film performance. While no model can account for all the cultural, economic, and competitive forces that influence a movie's success or failure, this study shows that methods driven by data can evidently support or replace intuition and experience in the film industry. By incorporating quantitative modeling with real-world scenarios in a creative industry, this research can help support the foundation for future work aimed at predicting accuracy and support more evidence based strategies in the film industry.

REFERENCES

[1] Gautam Kunapuli. Ensemble methods for machine learning. Manning Publications, 2023

[2] Yandex. CatBoost. 2026. url: https://catboost.ai