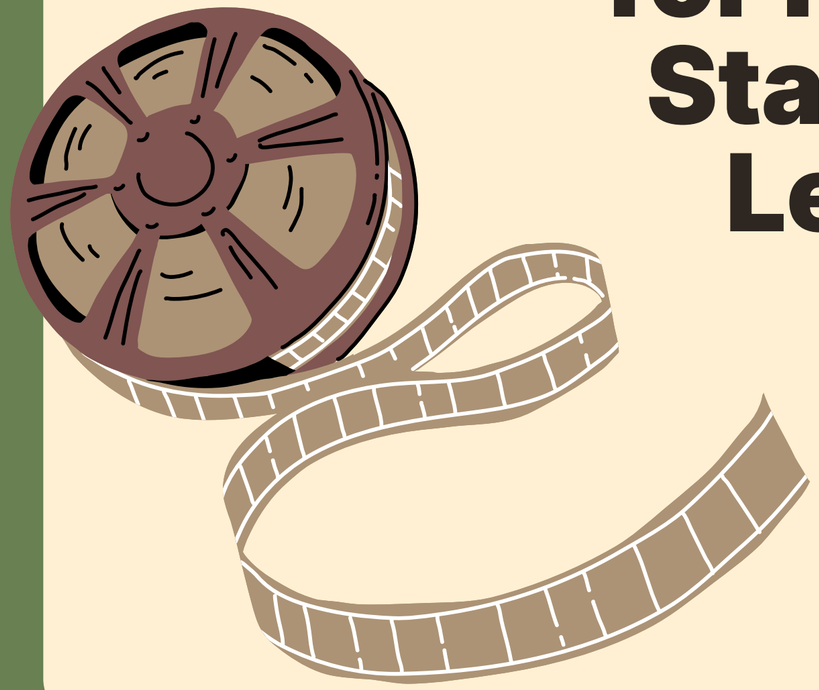
Deriving Predictive Models for Film Success Using Statistical & Machine Learning Methods

By Patricia Chen



Can we predict the success of a film before it is released?

Why is this important?

- Producing films is costly — producers want to profit
- It is hard to predict movie success due to the high variability in films
- Analyze trends and patterns that lead to high performing movies

How are we going to do this?

- Create two different predictive models
  - Risk scoring model
  - CatBoostRegressor model
- Determine the accuracy of models
- Find patterns and trends
- Test our models on new movies

Introduction

How Movies Were Invented 1870s-1900s

- Photography
- “The Horse in Motion”
- Kinetograph

Rise of Hollywood 1900s-1950s

- Thomas Edison
- Motion Picture Patents Company
- California migration

New Hollywood 1950s-1980s

- Fall of Hayes Code
- Movie Brats
- Blockbusters

Movies Today 1980s-present

- Technological advancements
  - Tapes
  - DVDs
  - CGI
  - Streaming

Previous Research

Lights, Camera, Profit
Alexandra L. Galbraith

- Logistic regression analysis
- Similar variables
- Further testing suggestions (awards)
- Breakeven response variable

Developing points-based risk-scoring systems in the presence of competing risks
Peter C. Austin et al.

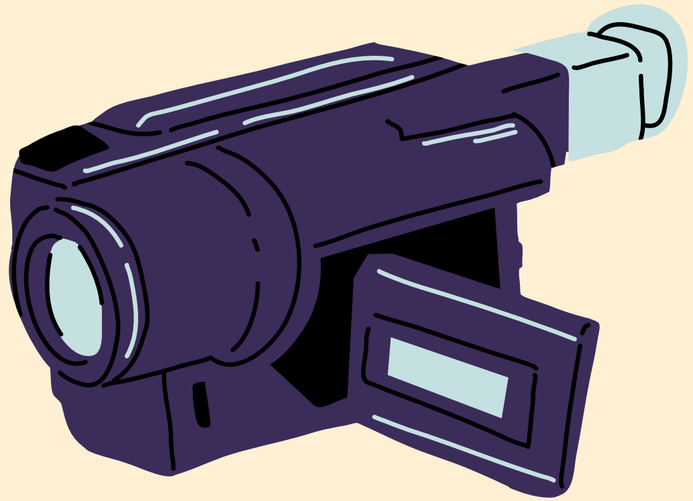
- In clinical context
- Step by step creation of risk scoring model

Predicting movie success based on pre-released features
Zulfiqar Ali Memon & Syed Muneeb Hussain

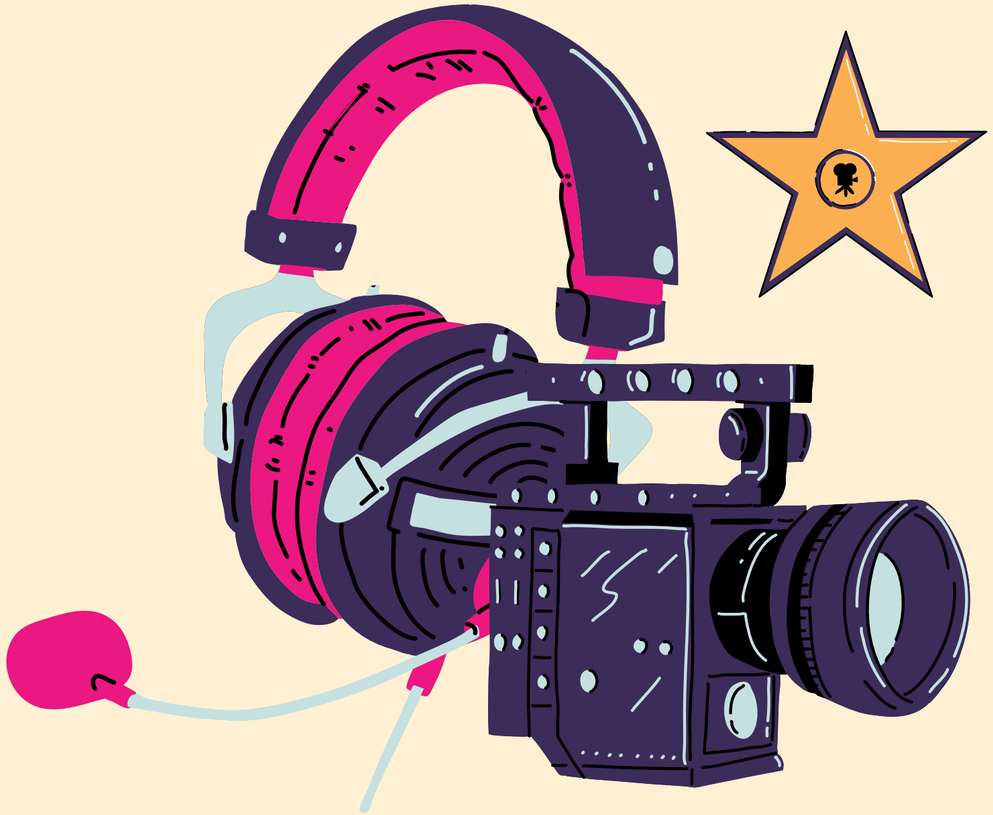
- Tested 11 machine learning models
- Concluded CatBoost model was most accurate

Data & Cleaning

- Datasets were collected from
  - IMDb
  - Rotten Tomatoes (Kaggle)
  - Metacritic (Kaggle)
  - Oscars (Kaggle)
  - Golden Globes (Kaggle)
  - Movie Budget/Revenue (Kaggle)
- 2,694 observations
- 40 variables
- Converted all predictors to categorical variables
- Response variables →



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



Risk Scoring Model

A risk scoring model is a structured framework that uses data, and weighted rules to assign numerical values to risks, quantifying their impact.

- Binary response variable
- Breakeven outcome (y,n)
- Logistic Regression
- Predict a movie’s probability of breaking even

Logistic Regression

Logistic Regression is a statistical technique used to predict the probability of a binary outcome based on one or more independent variables.
In finding the best model, we

- Compared AIC & residual deviance
- Used stepwise regression
- Added back in necessary variables

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

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Sig.
(Intercept)	0.69149	0.44118	1.57	0.11703	
Month2	-0.13964	0.25470	-0.55	0.58352	
Month3	-0.43677	0.24816	-1.76	0.07840	.
Month4	-0.38675	0.25673	-1.51	0.13195	
Month5	0.14567	0.25865	0.56	0.57330	
Month6	0.15853	0.24270	0.65	0.51364	
Month7	0.16577	0.24053	0.69	0.49070	
Month8	-0.53551	0.23927	-2.24	0.02522	*
Month9	-0.60894	0.24269	-2.51	0.01210	*
Month10	-0.56732	0.23572	-2.41	0.01609	*
Month11	-0.02254	0.23719	-0.10	0.92431	
Month12	0.04019	0.23675	0.17	0.86519	
RuntimeMedium	0.61826	0.14851	4.16	3.1e-05	***
RuntimeLong	1.12522	0.18745	6.00	1.9e-09	***
RatingNC17	-0.77300	1.00038	-0.77	0.43970	
RatingPG	-0.30155	0.31050	-0.97	0.33147	
RatingPG13	-0.45700	0.33552	-1.36	0.17318	
RatingR	-0.73353	0.33697	-2.18	0.02949	*
GG_Actor_History1_2	-0.24489	0.09719	-2.52	0.01174	*
GG_Actor_History3_plus	-0.12754	0.13847	-0.92	0.35701	
Oscars_Director_History1_2	-0.35210	0.17825	-1.98	0.04823	*
Oscars_Director_History3_plus	0.28815	0.37349	0.77	0.44040	
Distribution_CompanyHigh	0.34050	0.08686	3.92	8.9e-05	***
BudgetMedium	-0.61344	0.10768	-5.70	1.2e-08	***
BudgetHigh	-0.74557	0.13392	-5.57	2.6e-08	***
Action	-0.24631	0.13029	-1.89	0.05869	.
Adventure	-0.00495	0.13623	-0.04	0.97101	
Animation	0.58436	0.24574	2.38	0.01741	*
Biography	-0.26276	0.18038	-1.46	0.14520	
Comedy	-0.05830	0.12810	-0.46	0.64904	
Crime	-0.42902	0.12759	-3.36	0.00077	***
Documentary	0.48514	0.53263	0.91	0.36238	
Drama	-0.41324	0.11899	-3.47	0.00051	***
Family	-0.55689	0.21735	-2.56	0.01040	*
Fantasy	-0.45869	0.16754	-2.74	0.00619	**
History	-1.21627	0.30192	-4.03	5.6e-05	***
Horror	0.41966	0.17280	2.43	0.01516	*
Music	-0.07020	0.22540	-0.31	0.75546	
Musical	-1.06437	0.57324	-1.86	0.06335	.
Mystery	0.19642	0.15520	1.27	0.20566	
Romance	-0.09137	0.13069	-0.70	0.48445	
Sci-Fi	-0.30826	0.16916	-1.82	0.06842	.
Sport	-0.75671	0.29375	-2.58	0.00999	**
Thriller	0.14267	0.14058	1.01	0.31019	
War	-0.53306	0.35228	-1.51	0.13024	
Western	0.04763	0.51105	0.09	0.92575	

Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Null deviance: 3702.5 (df = 2693)
Residual deviance: 3426.5 (df = 2648)
AIC: 3518.5

Table 4.4: Final Logistic Regression Model

Risk Scoring Model

- Created a matrix where each row represents a movie and each column was a predictor (including the categories), with a value of 0 or 1
- Made a vector with the coefficients multiplied by 100
- Multiplied this matrix and vector to generate the total additive score of all predictor contributions

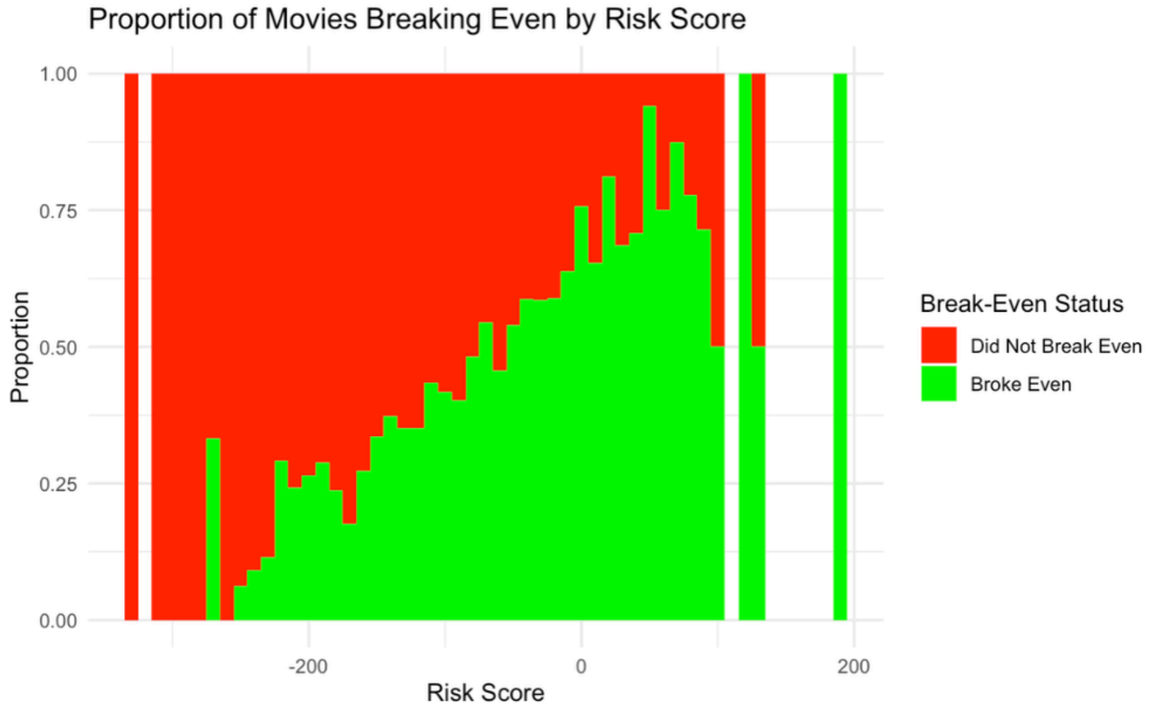
$$M\vec{v} = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1k} \\ m_{21} & m_{22} & \cdots & m_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ m_{N1} & m_{N2} & \cdots & m_{Nk} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{bmatrix} = \vec{s}.$$

Strata

- Positive correlation between risk score and breakeven proportion
- Divided risk scores into 5 equal strata

Strata	Number of Films	Mean Breakeven Proportion
1	539	0.230
2	539	0.356
3	539	0.432
4	539	0.536
5	538	0.673

Table 5.1: Risk Score Strata and Breakeven Proportion



Strata

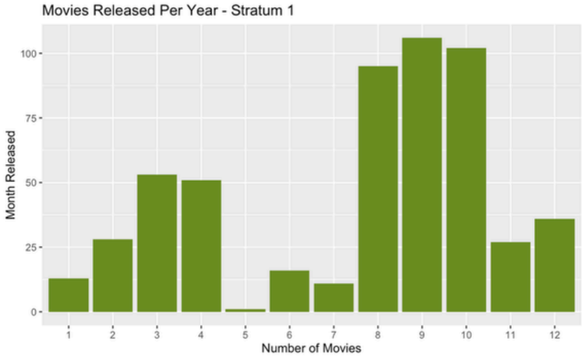


Figure 5.2: Distribution of movies released per month in Stratum 1

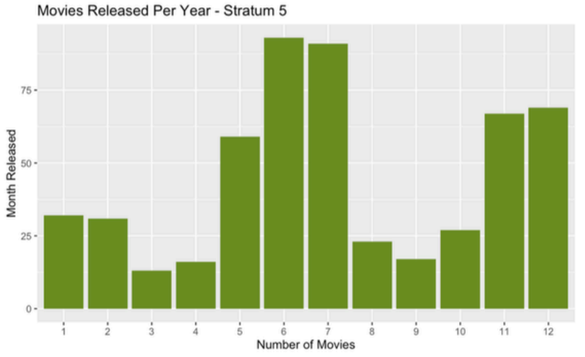


Figure 5.3: Distribution of movies released per month in Stratum 5

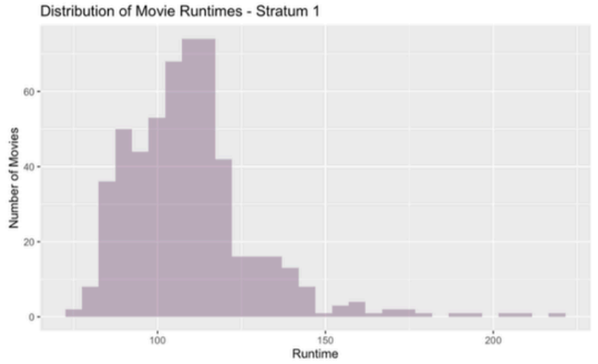


Figure 5.4: Distribution of movie runtimes in Stratum 1

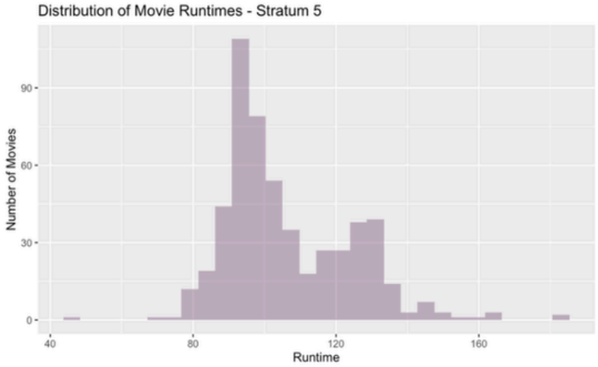


Figure 5.5: Distribution of movie runtimes in Stratum 5

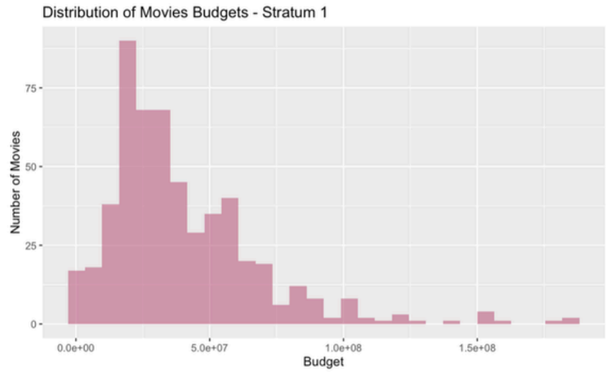


Figure 5.6: Distribution of movie budgets in Stratum 1

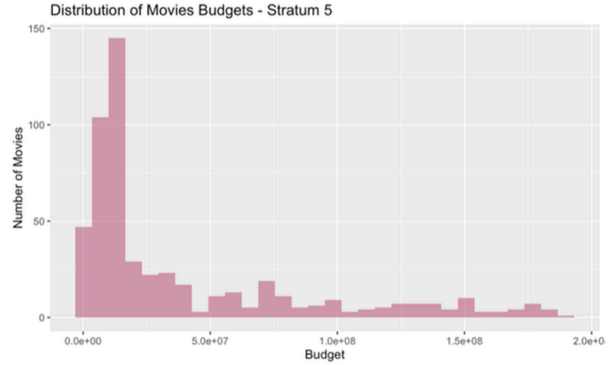


Figure 5.7: Distribution of movie budgets in Stratum 5

Strata

Genre

Stratum 1

Drama, Crime, Action, Comedy, and Biography

Stratum 5

Comedy, Adventure, Horror, Drama, and Thriller

MPAA

Stratum 1

313 R, 169 PG-13, 54 PG, 2 G

Stratum 5

192 R, 170 pg-13, 145 PG, 31 G

Distribution Company

Stratum 1

177 top 10, 362 not

Stratum 5

305 top, 233 not

CatBoostRegressor Model

CatBoost is a machine-learning algorithm that uses gradient boosting and decision trees for solving regression tasks and is designed to handle categorical variables well.

- Numerical response variable
- Revenue / Budget
- Gradient boosting
- Predict a movie’s proportional revenue

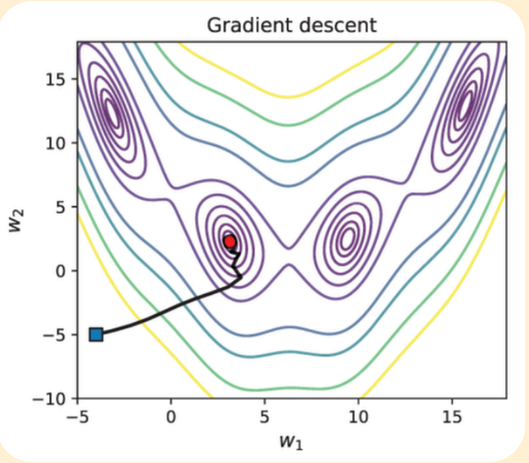
Gradient Descent

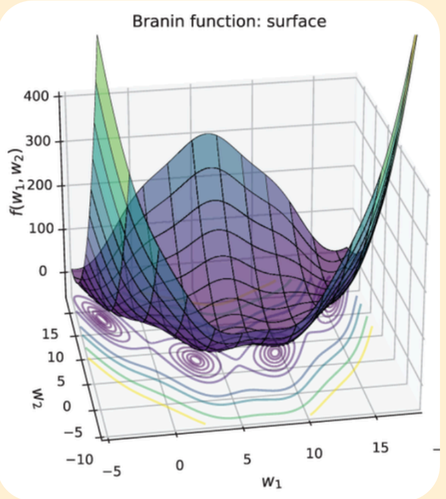
Gradient descent is an optimization method that involves minimizing a loss function.

- Loss function represents the difference between a model's predicted values and the actual values.
- The gradient represents the steepest ascent of a function
- The negative gradient of the loss function was taken to find the fastest way to the minimum.

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

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





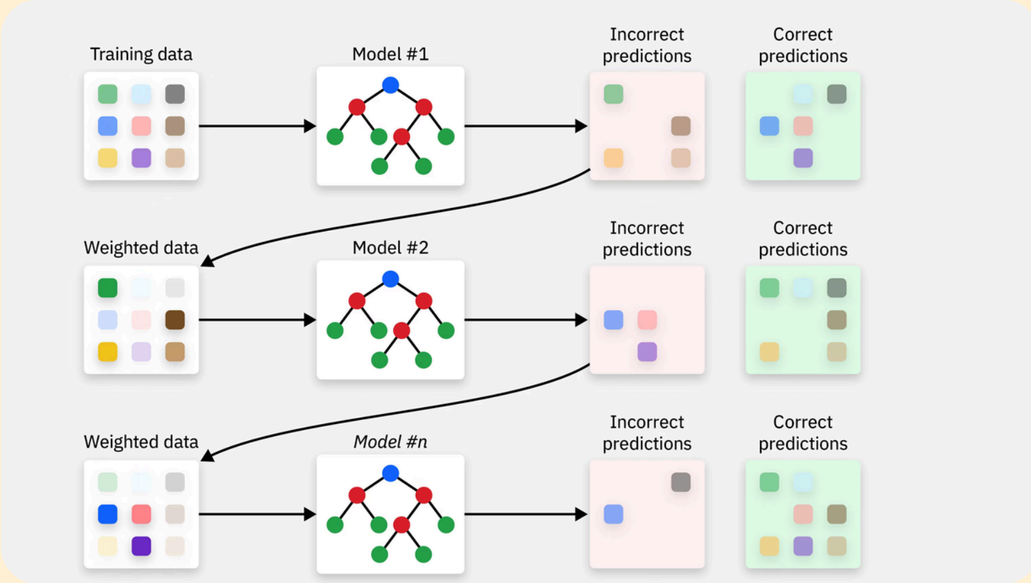
[1]

Gradient Boosting

Gradient boosting is a machine learning algorithm that builds models sequentially, with each new model (decision tree) trained to improve itself from the mistakes of the previous ones.

$$F_t(x) = F_{t-1}(x) + \alpha h_t(x),$$

$$F_T(x) = F_0(x) + \alpha \sum_{t=1}^{T} h_t(x)$$



[2]

CatBoostRegressor Model

- Trained model on log(Proportional Revenue + 1)
- Split the training and testing data 80/20 respectively
- Model ran through 1,000 sequential decision trees
- After predictions were generated, the values were converted back into the original scale

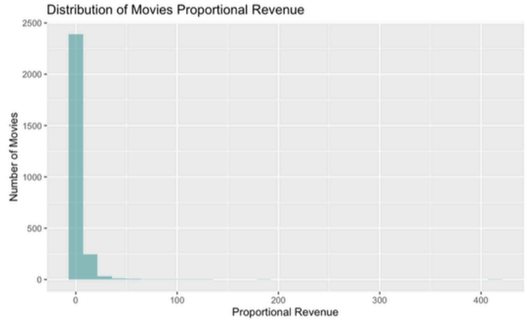


Figure 4.4: Distribution of proportional revenue before transformation.

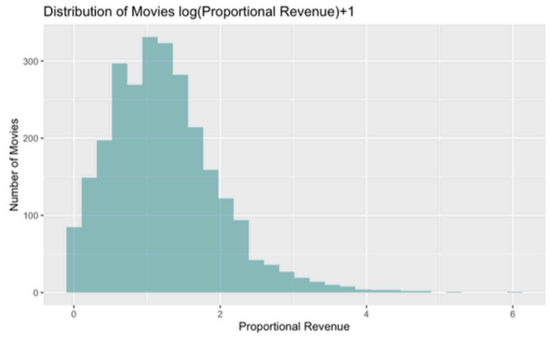


Figure 4.5: Distribution of proportional revenue after taking the natural log.

Results

- Errors are well below the data's standard deviation, despite the low R^2.
- Feature Importance shows a significance value for how important an independent variable is in generating the predictions [3].

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
Drama	2.7859
Crime	2.5166
Romance	1.7988
Action	1.4771
Family	1.3787
Thriller	1.2640
Sci-Fi	1.0990
Adventure	1.0264
History	0.9332
Biography	0.9157
Fantasy	0.8521
Animation	0.8059
Music	0.8003
Documentary	0.6331
War	0.6250
Musical	0.2816
Western	0.1873

Table 5.4: CatBoost Feature Importance for Predicting Proportional Revenue

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

Table 5.3: CatBoost Results

Testing On New Films

Additionally, we decided to test the accuracy of our two models on newer films not included in our data.

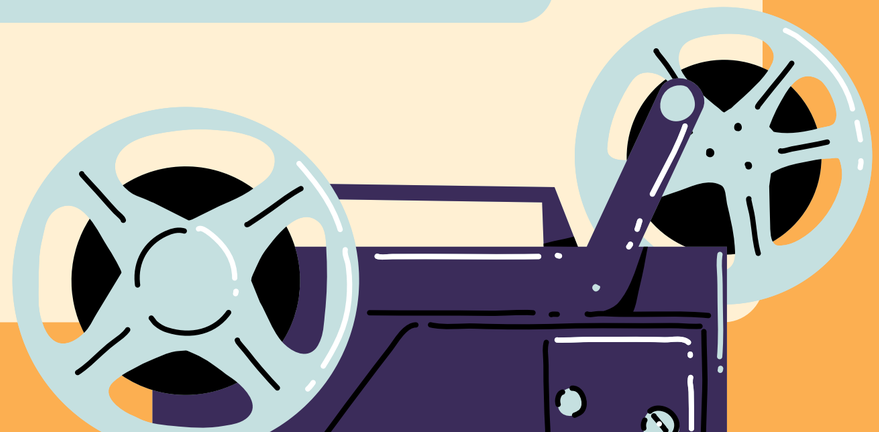
Film	Release Year	Pred. Breakeven Prop.	Breakeven	Pred. Prop. Rev.	Prop. Rev.
Spiderman 3	2022	0.31	Yes	2.76	9.61
Top Gun	2022	0.50	Yes	2.85	8.76
Barbie	2023	0.30	Yes	3.21	9.65
Challengers	2024	0.00	No	1.06	1.74
Sinners	2025	0.56	Yes	2.09	3.70

Table 5.5: Models' Performance on Post-2020 Films

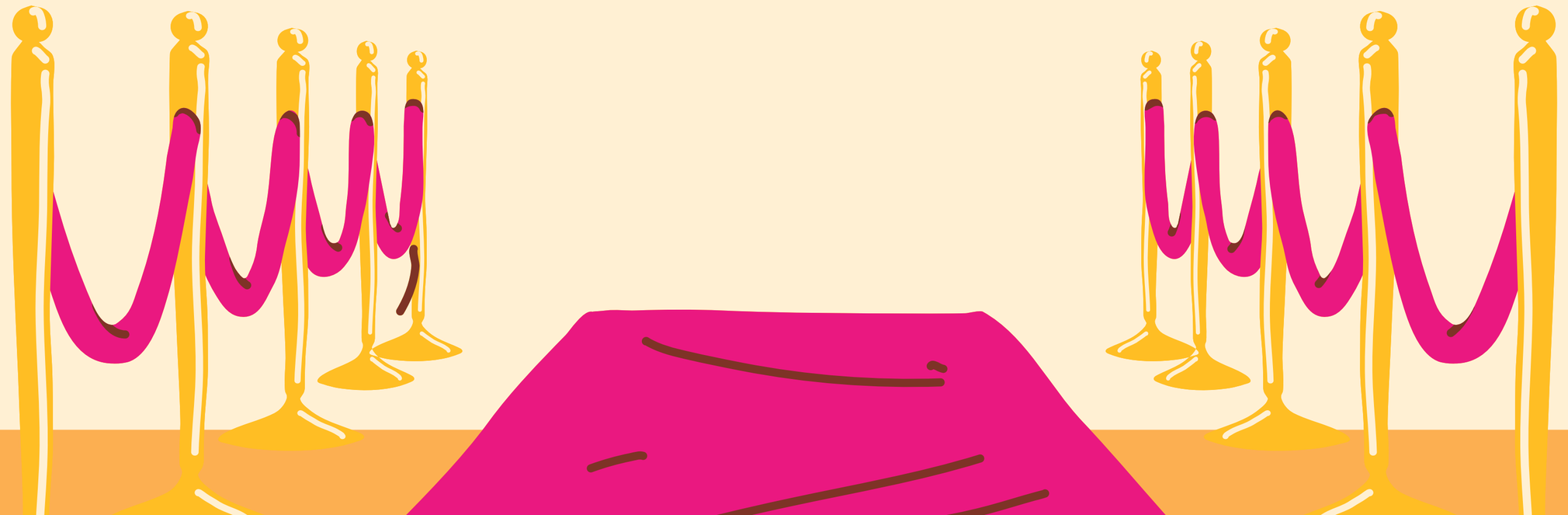
Limitations & Further Research

- Data was gathered from Kaggle
- Lack of randomness
- Risk scoring model overfitting
- Range of movies 1970-2020
- Double counting with genres
- Under-representation in G-rated and PG-rated movies and musicals
- Only uses categorical variables
- Unpredictable nature of film performance

- Explore other target variables
- More award variables
- Sequel or franchise variables
- Social media & audience engagement
- Using continuous variables
- Expanding dataset



Thank you!
Any questions?



References

[1] Gautam Kunapuli. Ensemble methods for machine learning. Manning Publications, 2023
[2] IBM. What Is Gradient Boosting? IBM. https://www.ibm.com/think/topics/gradient-boosting
[3] Yandex. CatBoost. 2026. url: https://catboost.ai