READY CLUSTER ONE: OPTIMIZING FILM SUCCESS WITH DATA SCIENCE

Joseph Richardson & Mohsin Md Abdul Karim *with* Dr. Bin Luo

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

INTRODUCTION

Imagine you're a movie producer tasked with deciding how much to invest in an unproduced screenplay. You know the screenplay's subgenre and genre, and you have an estimated runtime based on its page count. How do you know what budget tier to assign the movie? Should you invest in a smaller, more risk-averse budget, or bet big on a tentpole film?

Category	Lower Bound (in millions)	Upper Bound (in millions)	Number of Movies
Low	2	20	1,135
Medium	20	50	1,480
High	50	100	1,060
Tentpole	100	>100	903

Table 1: Budget Tier Distribution

In this poster, we help answer that question by analyzing historical movie data and building a model that suggests an optimal budget tier for any given film based on its features.

Using past data from a wide variety of films, we applied a Finite Mixture Model (FMM) to identify clusters of films with similar characteristics (subgenre, genre, and runtime). These clusters were then used to predict the most likely budget tier for a new screenplay—balancing profit potential with risk.

METHODS

DATA SOURCES AND FEATURES

- IMDb Dataset: Includes movie genre, subgenre, release year, and user ratings.
- TMDB Dataset: Movie profit, budget, and cast/crew information.
- Used page length as a proxy for movie runtime.
- Financial values adjusted for inflation (2023 USD).

CLUSTERING

- Applied FMM on stratified budget tiers to predict movie clusters.
- *K*-tuning performed using AIC, BIC, and cluster size analysis.
- Five-fold cross-validation used to evaluate FMM, regression, and random forest models.

MODEL EVALUATION

- Models evaluated using MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and CI (Confidence Interval) coverage for profit predictions.

RISK ADJUSTMENT

- Balanced ROI and profit for each test movie; $\lambda = 0.96$ penalized high-variance profits to favor stable returns.

DESCRIPTIVE ANALYSIS

GRAPH STRUCTURE & CAST/CREW RECOMMENDATIONS

- The graph is a *k*-partite network, providing a holistic view of movie production relationships across genres, subgenres, companies, cast, and crew.
- Using the graph, we identify the most frequent cast and crew members within each movie cluster, enabling recommendations for cast and crew selections that align with the historical success patterns of similar films.

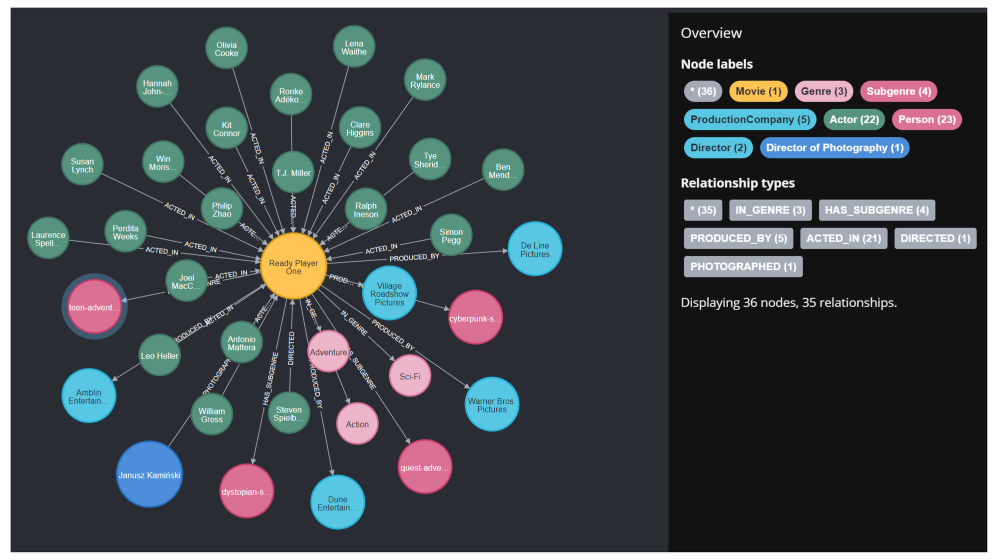

Figure 1: The sub-graph of Ready Player One

BASELINE MODELS

- To benchmark the FMM approach, we first trained two baseline models: linear regression and random forest.
- Both models used the same movie features (genre, subgenre, runtime) to predict profit directly.
- Their performance was compared against FMM in terms of MAE, RMSE, and R^2. Results are summarized in Table 2.

FMM MODEL SELECTION

- Although random forests achieved lower MAE and continuous-budget FMM had slightly better RMSE and R^2, the tiered FMM was chosen for consistency with graph-based cast and crew recommendations.
- Clustering was necessary to query the graph, and budget-tier FMM provided better CI coverage and aligned naturally with producer budget categories.

Model	Budget Type	MAE	RMSE	CI Coverage
Linear Regression	Continuous Budget	$161,749,779	$272,804,247	N/A
Random Forest	Continuous Budget	$144,066,187	$289,704,096	N/A
FMM	Continuous Budget	$155,327,500	$248,644,842	6.06%
FMM	Budget Tiers (K=23/20/25/20)	$158,133,237	$254,392,057	14.29%

Table 2: Model Performance Comparison, $\lambda = 0.96$

BUDGET OPTIMIZATION

BUDGET TIER RECOMMENDATIONS

- Tentpole films offer higher expected profits but also much higher variance, making them riskier investments.
- Low-budget films provide more stable but smaller profits, appealing to producers seeking lower financial risk.

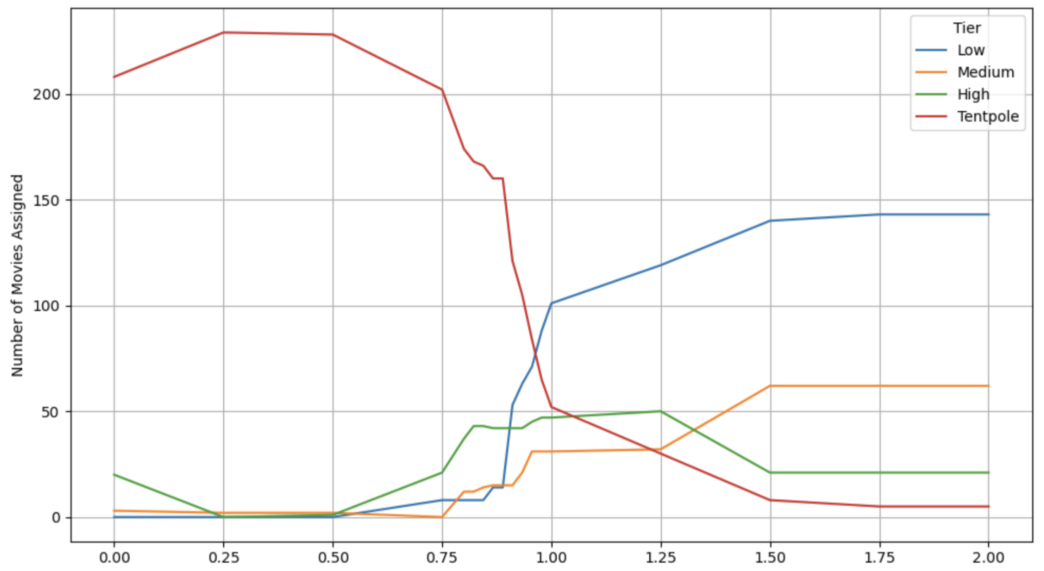

Figure 2: Risk Penalty by Budget Tier

PROFIT AND RISK TRADE-OFF

- Risk adjustment penalized movies with high variance, typically tentpoles, making low-budget films more attractive for certain screenplays, especially when profit uncertainty was high.
- This allowed the model to recommend budget tiers that balance expected returns with downside risk while avoiding extreme outcomes.

$$U(x) = \text{profit}(x) - \lambda \cdot \sigma_{\text{profit}}(x)$$

- The lambda value (λ) can be tuned to match a producer's risk tolerance; we selected $\lambda = 0.96$ to create a conservative model that balances tentpole opportunities with low-budget stability and reduces exposure to high-variance investments.

RESULTS

- We evaluated four real-world test screenplays — *Love Actually*, *Talladega Nights*, *Scarface*, and *Nope* — using our risk-adjusted cluster recommendation system.

Movie	Actual Tier	Suggested Tier	Actual Profit	Predicted Profit	Outcome
Love Actually	High	High	$574M	$144M	Conservative
Talladega Nights	Tentpole	Medium	$207M	$52M	Conservative
Scarface	High	High	$384M	$298M	Accurate
Nope	High	Tentpole	$111M	$401M	Optimistic

Table 3: Budget tier predictions and profit outcomes for test screenplays, $\lambda = 0.96$.

- Overall, the model tended to recommend more conservative budget tiers, particularly for movies like *Love Actually* and *Talladega Nights*, which achieved high profits but carried significant financial risk.
- The FMM approach successfully matched or reasonably downgraded budgets for three out of the four test cases, prioritizing financial stability over aggressive investment strategies.
- The model's recommendations are sensitive to the chosen lambda value ($\lambda = 0.96$); a lower lambda would have favored more tentpole investments with higher predicted profits, while a higher lambda would have pushed the model toward even safer, lower-budget recommendations.

CONCLUSIONS

- Finite mixture models effectively predict screenplay revenue and recommend budget tiers by clustering similar movies and balancing risk with a tunable penalty ($\lambda = 0.96$).
- While profit prediction remains noisy, especially for tentpole films, risk-adjusted estimates outperform direct regression methods. The model suggested conservative budget tiers for several test movies, demonstrating better investment guidance than baseline models.

FUTURE WORK

- We plan to refine budget recommendations by modeling profit distributions within FMM clusters to reduce noise and improve predictive precision.
- Future experiments will explore dynamically tuning the lambda value based on screenplay features to better balance risk and reward.
- To address the high variability within clusters, we also plan to test alternative clustering strategies or post-clustering adjustments to stabilize profit and ROI predictions.