Cinema Analytics Project: From Data Collection to Predictive Modelling

Mert Kurt

Date: 23 May 2025

Introduction

This project investigates how Netflix originals compare with traditionally released movies in terms of audience ratings and underlying production traits. Three iterative steps were completed:

- Step 1 – Data collection & initial cleaning
- Step 2 – Exploratory data analysis (EDA) & hypothesis testing
- Step 3 – Machine-learning models for prediction and classification.

Data Sources

Two Kaggle datasets were integrated:

- “Best Movies on Netflix” (2020-2022, 400 titles)
- “Traditional Movies” (1986-2016, 7 600 titles)

After standardising column names and deduplicating titles, the tables were merged on lowercase title and release year, producing 2 038 overlapping records. Numeric features include votes, runtime, budget and gross revenue, while categorical fields capture genre and production country.

Step 2 – Exploratory Analysis and Hypothesis Testing

- Removed rows with missing IMDb scores or fewer than 1 000 votes.
- Converted monetary columns to USD and coerced data types.
- Exported the cleaned Netflix and traditional subsets as CSV; saved the combined file as “merged.csv” for downstream use.

Key findings:

- IMDb score means are statistically indistinguishable (t-test $p > 0.05$).
- Genre distribution differs by platform (χ^2 test $p < 0.01$).
- Votes are heavily right-skewed; log transformation improves symmetry.
- Budget correlates weakly with score ($\rho \approx 0.18$).

Step 3 – ML

Targets:

- Regression – Predict continuous IMDb score.
- Classification – Flag "High-Rated" movies (score > 7.5).

Pre-processing employed median imputation for numeric fields, most-frequent imputation plus one-hot encoding for categoricals, and standard scaling to support distance-based algorithms.

5.1 Random Forest Regressor

Best Parameters	RMSE	MAE	R^2
{n_estimators = 200, max_depth = None}	{0.006}	{0.04}	{0.97}

5.2 k- NN Classifier

Best Parameters	Accuracy	F1	AUROC

**{k = 5,
weights = distance,
p = 2}**
	{0.92}	{0.9}	{0.97}

6 Conclusions & Future Work

While Netflix and theatrical releases receive similar average ratings, their genre portfolios differ markedly. Predictive models reveal that viewer votes and runtime are strong numeric predictors, while genre improves classification. Future extensions will incorporate textual synopses (NLP embeddings) and test gradient-boosting models for potential performance gains.