# Domain Understanding - Machine Learning Perspective

**Movie Success Prediction Project**

By Anton Horvat

Contents

[Domain Understanding - Machine Learning Perspective 1](#_Toc208312092)

[Problem Formulation 3](#_Toc208312093)

[Target Variable Definition 3](#_Toc208312094)

[Dataset Selection 4](#_Toc208312095)

[Feature Engineering Opportunities 5](#_Toc208312096)

[Model Considerations 5](#_Toc208312097)

[Evaluation Strategy 6](#_Toc208312098)

[Expected Challenges 6](#_Toc208312099)

[Success Criteria 7](#_Toc208312100)

[References 8](#_Toc208312101)

# Problem Formulation

**Machine Learning Problem Type:** Multi-class Classification

- **Task:** Predict movie financial success category before release

- **Target Variable:** Three-class categorical outcome (Hit/Break-even/Flop)

- **Learning Type:** Supervised learning with labeled historical data

# Target Variable Definition

**Profitability Classification** based on industry financial standards:

- **Flop:** Revenue < 1x Production Budget ([immediate financial loss, considered industry failure](https://www.hollywoodreporter.com/business/business-news/why-film-budgets-are-blown-1135112/))

- **Break-even:** Revenue = 2-2.5x Production Budget ([covers marketing and distribution costs](https://stephenfollows.com/how-movies-make-money-hollywood-blockbusters/))

- **Hit:** Revenue > 2.5x Production Budget ([profitable after all industry costs](https://entertainment.howstuffworks.com/movie-cost.htm))

**Industry Standard Rationale:** [Production budget represents only filming costs](https://www.investopedia.com/articles/investing/093015/how-exactly-do-movies-make-money.asp). Total investment includes marketing expenditure (typically 50-100% of production budget) and distribution costs, while [theaters retain approximately 50% of box office revenue](https://www.thebalancecareers.com/how-do-movie-theaters-make-money-1793856).

**ML Rationale:** Classification chosen over regression because:

- Business decisions require categorical outcomes (invest/don’t invest)

- Inherent uncertainty in exact revenue prediction makes ranges more reliable

- Industry stakeholders think in risk categories, not precise numbers

# Dataset Selection

**Primary Source:** [The Movies Dataset (Kaggle)](https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset)

- Real-time access to comprehensive movie metadata with over 1 million movies

- Volume: Access to extensive movie database with up-to-date information

- Format: JSON API responses for flexible data collection and integration

Key Advantages:

* Includes budget and revenue data essential for profitability analysis
* Real-time data updates ensuring current information
* Comprehensive metadata including cast, crew, genres, and financial performance
* Flexible API endpoints for targeted data collection

**Alternative Datasets Considered:**

* [TMDB 5000 Movie Dataset (Kaggle)](https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata) - Static snapshot, limited to older data
* [Full TMDB Movies Dataset 2024 (Kaggle)](https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies) - Large but static dataset

**Why TMDB API Over Static Datasets:**

* **Real-time Access:** Current movie information and recent releases
* **Data Freshness:** Up-to-date financial performance and metadata
* **Comprehensive Coverage:** Access to full TMDB database, not limited subsets
* **API Flexibility:** Custom queries for specific data requirements
* **Professional Development:** Experience with industry-standard API integration

**Comparison with Academic Research:**

- [Sharda & Delen (2006)](https://www.researchgate.net/publication/222530390_Predicting_box-office_success_of_motion_pictures_with_neural_networks) used smaller datasets with limited feature sets and achieved 75% accuracy

- [Lash & Zhao (2016)](https://www.tandfonline.com/doi/abs/10.1080/07421222.2016.1243969) focused on social network data, not comprehensive movie features

- Modern [Kaggle movie datasets](https://www.kaggle.com/search?q=movie+dataset) provide larger, more comprehensive datasets

**Dataset Features:**

- **Predictive Features:** Budget, genre, cast, director, release date, runtime, production company, rating

- **Target Labels:** Box office revenue for profitability calculation

# Feature Engineering Opportunities

**Categorical Features:**

- Genre (multi-hot encoding for multiple genres)

- Content rating (ordinal encoding: G < PG < PG-13 < R)

- Release season (winter/spring/summer/fall)

**Numerical Features:**

- Production budget (log transformation for skewed distribution)

- Runtime (standardization)

- Release year (temporal trends)

**Derived Features:** - Star power index (based on cast’s previous box office performance)

- Director success rate (historical hit percentage)

- Competition level (number of similar releases in same period)

# Model Considerations

**Algorithm Suitability:**

- **Nearest Neighbors:** Naturally interpretable (“similar movies performed like this”)

- **Decision Trees:** Provide clear decision rules

- **Random Forest:** Handle mixed data types and feature interactions

- **Logistic Regression:** Baseline for comparison

**Literature-Based Algorithm Selection:**

- [Sharda & Delen (2006)](https://www.academia.edu/17161024/Predicting_box-office_success_of_motion_pictures_with_neural_networks) achieved 75% accuracy with neural networks but lacked interpretability

- Academic surveys show ensemble methods perform well for movie prediction tasks

- Focus on explainable models over pure accuracy optimization

# Evaluation Strategy

**Business-Aligned Metrics:**

- **Precision for “Hit” class:** Minimizing false positive investments

- **Recall for “Flop” class:** Identifying potential failures

- **Overall accuracy:** Balanced performance across all categories

**Cross-Validation:** Time-based splits to prevent data leakage (train on older movies, test on newer)

# Expected Challenges

**Data Quality Issues:**

- Missing budget information for independent films

- Inflation adjustment needed for historical comparisons

- International vs domestic revenue considerations

**Feature Complexity:**

- Star power quantification requires aggregation of cast data

- Genre interactions (action-comedy vs pure action)

- Seasonal release effects vary by genre

**Model Interpretability:**

- Need to explain predictions to non-technical stakeholders

- Feature importance must align with industry intuition

- Avoid black-box models that can’t justify investment decisions

## 

# Success Criteria

**Technical Performance:**

- Classification accuracy > 65% (better than random 33%)

- Balanced performance across all three classes

- Feature importance aligns with domain knowledge

**Business Value:**

- Predictions are interpretable and actionable

- Model identifies key risk factors for investment decisions

- Performance generalizes to unseen movies (proper validation)

# Table for ML Understanding:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ML Approach** | Description of Approach | Target Variable | Feature Selection | **Why This Approach** |
| k-Nearest Neighbors (k-NN) | Distance-based algorithm that predicts movie success by finding k most similar movies in the training set and using their success categories to make predictions. Uses scaled features to calculate Euclidean distance between movies. | Success category encoded as numeric  **(**  **0=Flop,**  **1=Break-even,**  **2=Hit)**  using LabelEncoder | 4 features:  **budget\_log (financial investment),**  **runtime (production quality),**  **vote\_average (audience appeal),**  **imdb\_rating (critical reception)** | **Interpretability:**  Naturally explainable to stakeholders - "this movie is similar to these 20 movies that were hits"  **Domain alignment:** Industry thinks in terms of comparable movies  **Iris k-NN lesson:**  Distance-based algorithms work well with scaled numeric features  **Business value:** Can show which historical movies influenced each prediction |
| **Random Forest** (future) | Ensemble of decision trees that makes predictions by majority voting. Each tree learns different patterns from random subsets of data and features. | Same: success\_encoded (0, 1, 2) | Same 4 features with potential kernel transformations (RBF, polynomial) | **Clear decision boundaries:** Works well when classes are separable  **Hyperparameter impact:** SVM assignment showed C values dramatically affect performance (57% → 82%)  **Handles imbalanced data:** Can use class weights to address Hit class bias  **Non-linear patterns:** RBF kernel can capture complex budget-rating-success relationships |
| **Support Vector Machine (SVM)** (future) | Creates decision boundaries that maximize separation between success categories in high-dimensional space. Can use different kernels for non-linear patterns. | Same: success\_encoded (0, 1, 2) | Same 4 features with potential kernel transformations (RBF, polynomial) | **Clear decision boundaries:** Works well when classes are separable  **Hyperparameter impact:** SVM assignment showed C values dramatically affect performance (57% → 82%)  **Handles imbalanced data:** Can use class weights to address Hit class bias  **Non-linear patterns:** RBF kernel can capture complex budget-rating-success relationships |
| **Logistic Regression** (future) | Statistical model that predicts probability of each success category using weighted combination of features. Provides coefficients showing each feature's contribution. | Same: success\_encoded (0, 1, 2) using multinomial logistic regression | Same 4 features, potentially with polynomial terms or interaction features | **Baseline comparison:** Simple model to benchmark against complex algorithms  **Interpretable coefficients:** Shows exactly how budget/rating increase affects success probability  **Probability outputs:** Gives confidence scores (80% chance of Hit) not just predictions  **Wine assignment lesson:** Starting simple before adding complexity helps understand feature relationships |

# References

**Academic Literature:**

- [Lash, M. T., & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability.](https://www.tandfonline.com/doi/abs/10.1080/07421222.2016.1243969) *Journal of Management Information Systems*, 33(3), 874-903.

- [Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks.](https://www.researchgate.net/publication/222530390_Predicting_box-office_success_of_motion_pictures_with_neural_networks) *Expert Systems with Applications*, 30(2), 243-254.

**Industry Sources:**

- [Box Office Mojo. (2024). Movie Revenue Database.](https://www.boxofficemojo.com/)

- [Follows, S. (2024). How Movies Make Money: Hollywood Blockbusters.](https://stephenfollows.com/how-movies-make-money-hollywood-blockbusters/)

- [Investopedia. (2024). How Exactly Do Movies Make Money?](https://www.investopedia.com/articles/investing/093015/how-exactly-do-movies-make-money.asp)

**Data Sources:**

- [The Movies Dataset - Kaggle](https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset)

- [IMDB Non-Commercial Datasets](https://developer.imdb.com/non-commercial-datasets/)

- [Statista. (2024). North American Box Office Statistics.](https://www.statista.com/statistics/187069/north-american-box-office-gross-revenue-since-1980/)