# Movie Success Prediction Project Proposal

**Student Name:** Anton Horvat  
**Course:** Artificial Intelligence, Machine Learning & Data  
**Semester:** 4  
**Date:** 08-09-2025

Contents

[Movie Success Prediction Project Proposal 1](#_Toc208240886)

[1. What are you going to do? 3](#_Toc208240887)

[2. Why are you going to do that? 3](#_Toc208240888)

[2.1 Domain Understanding Research 3](#_Toc208240889)

[Research Question 3](#_Toc208240890)

[Research Methods 3](#_Toc208240891)

[Domain Research Findings 4](#_Toc208240892)

[Connection to Technical Approach 5](#_Toc208240893)

[2.2 Analytic Approach 5](#_Toc208240894)

[Problem Type Classification 5](#_Toc208240895)

[Target Variable Definition 5](#_Toc208240896)

[Algorithm Type Justification 6](#_Toc208240897)

[2.3 Data Requirements 6](#_Toc208240898)

[Objectives 6](#_Toc208240899)

[Data Characteristics 6](#_Toc208240900)

[Data Sources 6](#_Toc208240901)

[Expected Good Indicators 8](#_Toc208240902)

[2.4 Why This Project Is Still Valuable Despite Existing Movie Prediction Models 8](#_Toc208240903)

[Educational vs. Commercial Purpose 8](#_Toc208240904)

[Direct Teacher Benefits 9](#_Toc208240905)

[3. Who is involved (stakeholders)? 10](#_Toc208240906)

[4. When is it happening (planning, iterations, Gantt chart)? 11](#_Toc208240907)

[Weeks 1-3: Foundation & Data Setup 11](#_Toc208240908)

[Weeks 4-6: Core Model Development 11](#_Toc208240909)

[Weeks 7-8: Optimization & Business Application 11](#_Toc208240910)

[Weeks 9-10: Documentation & Portfolio 11](#_Toc208240911)

[5. How are you going to do it (inferencing, delivery)? 12](#_Toc208240912)

[Following IBM Data Science Methodology: 12](#_Toc208240913)

[Iteration Zero Constraints (Quick & Dirty Proof of Concept): 12](#_Toc208240914)

[Competency Development Through Methodology: 13](#_Toc208240915)

[Technical Implications 13](#_Toc208240916)

[Iterative Process: 13](#_Toc208240917)

[Why This Project is Ideal for Competency Development 14](#_Toc208240918)

[References 15](#_Toc208240919)

[Literature Study Sources 15](#_Toc208240920)

[Industry Analysis Sources 15](#_Toc208240921)

[Competitive Analysis Sources 15](#_Toc208240922)

[Industry Success Metrics Sources 15](#_Toc208240923)

[Statistical Data Sources 15](#_Toc208240924)

## 1. What are you going to do?

I will develop a machine learning system that predicts movie success by classifying films into three profitability categories: Hit, Break-even, or Flop. The system will analyze pre-release factors like budget, genre, cast, director, and release timing to predict whether a movie will be financially successful.

**What makes this different:** I will focus heavily on making the AI explainable - ensuring that anyone, even someone with no AI knowledge, can understand why the model makes its predictions. For example, being able to explain “This movie will likely flop because it’s a $200M horror film releasing in February, and the model learned that expensive horror movies historically struggle in winter months.”

## 2. Why are you going to do that?

**Personal Motivation:** As someone who genuinely loves movies and follows the entertainment industry closely, I’ve always been fascinated by the unpredictable nature of box office success. Why do some films with massive budgets and A-list casts fail spectacularly, while others with modest resources become unexpected hits? This project lets me combine my passion for movies with systematic AI analysis.

**Real-World Problem:** The entertainment industry represents one of the most high-risk investment environments, with [approximately 6 out of 10 films failing to break even](https://www.statista.com/statistics/187069/north-american-box-office-gross-revenue-since-1980/). [The global film industry is worth over $100 billion annually](https://www.statista.com/outlook/dmo/media/movies/worldwide), yet investment decisions often rely on subjective factors rather than data-driven analysis.

**Why This Matters:**

**Economic Impact:** Billions of dollars are wasted annually on poor film investments. Better prediction could redirect resources toward more promising projects.

**Perfect Learning Opportunity:** This project allows me to demonstrate all five course competencies while working on something I’m genuinely passionate about, making the learning process more engaging and meaningful.

## 2.1 Domain Understanding Research

### Research Question

“What factors historically determine movie financial success, and how can these factors be quantified and utilized for predictive modeling in entertainment industry investment decisions?”

### Research Methods

Following emancipatory research principles to enhance my understanding of the movie industry domain, I employed multiple research methods from the HBO-i framework:

**Literature Study:** Analysis of academic papers on movie success prediction, industry reports from major studios, and box office analysis publications to understand established success factors.

**Sources Consulted:**

- [Sharda & Delen (2006) - “Predicting box-office success of motion pictures with neural networks”](https://www.sciencedirect.com/science/article/abs/pii/S0957417405001399)

- [Lash & Zhao (2016) - “Early predictions of movie success”](https://www.tandfonline.com/doi/abs/10.1080/07421222.2016.1243969)

**Competitive Analysis:** Review of existing movie prediction systems and academic research to understand current approaches, their limitations, and opportunities for improvement.

**Systems Analyzed:**

- [Kaggle TMDB Box Office Prediction Competition](https://www.kaggle.com/c/tmdb-box-office-prediction) methodologies and results

- [GitHub Movie Success Prediction Projects](https://github.com/search?q=movie+success+prediction) with documented approaches

- Academic implementations from [IEEE Xplore Digital Library](https://ieeexplore.ieee.org/search/searchresult.jsp?queryText=movie%20box%20office%20prediction)

**Document Analysis:** Examination of entertainment industry reports, box office databases, and financial analyses to identify patterns and quantifiable success metrics.

**Primary Sources:**

- [Box Office Mojo](https://www.boxofficemojo.com/)

- Industry standard for revenue tracking

- [The Numbers](https://www.the-numbers.com/) - Comprehensive financial database

### Domain Research Findings

**Historical Success Factors Identified:**

- **Budget allocation patterns**: Movies with budgets between $15-80 million demonstrate highest ROI success rates

- **Genre performance**: Action and comedy films demonstrate more predictable success patterns than horror or drama

- **Release timing**: Summer and holiday seasons significantly impact box office performance

- **Star power quantification**: Leading actor’s previous box office performance correlates with opening weekend success

- **Director track record**: Directors’ historical success rates are strong predictors of future performance - **Production company influence**: Major studio backing affects marketing reach and distribution success

**Industry Decision-Making Insights:** Research revealed that current industry practices rely heavily on subjective assessments, with quantitative analysis limited to basic budget-to-revenue comparisons. This creates an opportunity for more sophisticated predictive modeling.

**Target Variable Justification:** Based on domain research, profitability ratio (revenue/budget) emerges as the most meaningful success metric, as it accounts for investment risk and provides actionable business insights regardless of absolute revenue figures.

### Connection to Technical Approach

This domain understanding directly informed my technical decisions:

- **Target variable choice**: Profitability classification based on industry success thresholds

- **Feature selection**: Focus on pre-release factors actually used by industry professionals

- **Dataset requirements**: Need for comprehensive movie database with both predictive features and outcome data

## 2.2 Analytic Approach

### Problem Type Classification

This project addresses a **Classification problem** - specifically answering “Is this movie going to be a Hit, Break-even, or Flop?” This falls under the “Is this A, B, C, or D?” category from the algorithm selection framework.

### Target Variable Definition

**Target Variable**: Movie success classification based on real-world industry profitability standards:

**Real Hollywood Success Categories** (Industry Standard):

- **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). Therefore, a $100M production budget requires $250M+ revenue to achieve actual profitability.

### Algorithm Type Justification

Classification algorithms are most appropriate for this problem because:

1. **Actionable insights**: Stakeholders need categorical decisions (invest/pass) rather than exact revenue predictions

2. **Uncertainty management**: Classification handles the inherent unpredictability of exact box office figures

3. **Business alignment**: Industry decision-making processes naturally align with categorical risk assessment

4. **Interpretability**: Classification models provide clearer explanations for business stakeholders

## 2.3 Data Requirements

### Objectives

To build a movie success classification model, I require comprehensive pre-release movie data paired with corresponding box office outcome data, enabling supervised learning for predictive analytics.

### Data Characteristics

**Structured Data Requirements:**

- **Volume**: Minimum 5,000 movies for statistical significance, target 10,000+ for robust training

- **Time Range**: Movies from 1990-2024 to capture modern industry patterns while ensuring revenue data availability

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

- **Target labels**: Box office revenue figures for profitability calculation

### Data Sources

**Primary Source:** [The Movie Database (TMDB) API](https://www.themoviedb.org/documentation/api)

* Real-time access to comprehensive movie databases with budget, revenue, and performance metrics
* Professional API integration experience for industry-standard development
* Current data access without dependency on static dataset limitations
* Flexible data collection for specific research requirements- Offline work capability for consistent development

**API Advantages over Static Datasets:**

* **Real-time updates**: Access to current movie information and recent releases
* **Comprehensive coverage**: Full TMDB database access, not limited subsets
* **Professional development**: Experience with industry-standard API integration
* **Data freshness**: Up-to-date financial performance and metadata
* **Flexible queries**: Custom data collection for specific requirements

**API Implementation Strategy:**

* **Rate limiting management**: 40 requests per 10 seconds for free tier
* **Authentication handling**: API key registration and secure access
* **Data validation**: Robust error handling and response processing
* **Systematic collection**: Strategic data gathering across multiple years

**Data Source Comparison:**

* **Kaggle datasets**: Include budget, revenue, and comprehensive movie metrics
* **IMDb public CSV files**: Lack budget/revenue information (ratings, cast, crew only)
* **IMDb API**: Contains budget/revenue data but requires API access and has rate limits

**Data Legality and Ethics**

* **Legal compliance:** TMDB API is publicly available for research and educational use
* **Ethical considerations**: Using publicly available movie performance data raises no privacy concerns
* **Attribution**: Proper citation of data source in all project documentation

### Expected Good Indicators

Based on domain research findings, anticipated strong predictors include:

- **Budget-Related Features**: Production budget relative to genre expectations

- **Talent Indicators**: Director’s historical success rate, lead actor’s box office track record

- **Market Timing Factors**: Release season optimization, competition level during release window

- **Content Characteristics**: Genre classification, content rating impact, runtime optimization

## 2.4 Why This Project Is Still Valuable Despite Existing Movie Prediction Models

### Educational vs. Commercial Purpose

**This is NOT about creating the world’s first movie prediction system.** This is about demonstrating competency development through systematic methodology application. The value lies in the learning process, not in revolutionary innovation.

**Clear Educational Justification:**

**1. Learning Vehicle, Not Commercial Product** - Movie prediction provides an ideal domain for demonstrating AI competencies because everyone understands the business context - The goal is mastering IBM Data Science Methodology, not competing with Netflix’s recommendation algorithms

**2. Explainable AI Focus Missing from Existing Research**

- [Sharda & Delen (2006)](https://www.sciencedirect.com/science/article/abs/pii/S0957417405001399) used neural networks with 75% accuracy but provided no explanations

- [Lash & Zhao (2016)](https://www.tandfonline.com/doi/abs/10.1080/07421222.2016.1243969) achieved high accuracy through complex social network analysis, but results were black-box

- **My focus: Can a non-technical person (my girlfriend) understand WHY the model made each prediction?**

**3. Modern Industry Standards Application** - Previous research used outdated success categories or complex 9-class systems

- I’m using real industry standards: the 2.5x revenue multiplier that studios actually use for investment decisions

**4. Systematic Methodology Demonstration**

- Existing research focused on algorithmic innovation, not educational methodology

- My project demonstrates complete IBM Data Science Methodology from business understanding through deployment

### Direct Teacher Benefits

**Assessment Clarity**: Movies provide universally understood evaluation context

- you can easily judge if my explanations make sense and if my model insights are reasonable.

**Competency Verification**: Entertainment domain enables clear demonstration of explainable AI

- if I can make movie predictions understandable to anyone, I’ve mastered the competency.

**Practical Learning**: Better than abstract datasets because I can validate model reasoning against real-world knowledge.

## 3. Who is involved (stakeholders)?

**Academic Environment:**

- **Course Instructors**: Will evaluate my competency development across all five rubric areas

- **Peer Students**: Will provide feedback during presentations and collaborative learning

**Personal Support:**

- **My Girlfriend**: As a fellow movie enthusiast, she’s genuinely interested in this project. Her straightforward, no-nonsense feedback style is perfect for testing my **Explainable AI** competency - if I can make her understand why the model predicts certain outcomes (without any AI background), then I’ve truly mastered making AI transparent and understandable.

This combination ensures I get both technical feedback for competency development and real-world testing of my ability to communicate AI insights to non-technical audiences.

## 4. When is it happening (planning, iterations, Gantt chart)?

**8-10 Week Challenge Timeline:**

### Weeks 1-3: Foundation & Data Setup

* **Week 1**: Domain research completion and proposal finalization
* **Week 2**: IMDb dataset acquisition and initial data exploration
* **Week 3**: Data cleaning and preparation following AI methodology

### Weeks 4-6: Core Model Development

* **Week 4**: Iteration Zero - Nearest Neighbors baseline model
* **Week 5**: Multiple classification algorithms and comparison
* **Week 6**: Model evaluation and feature importance analysis

### Weeks 7-8: Optimization & Business Application

* **Week 7**: Model optimization and explainable AI development
* **Week 8**: Business recommendations and final presentation

### Weeks 9-10: Documentation & Portfolio

* **Week 9**: Complete technical documentation and code review
* **Week 10**: Final portfolio compilation and challenge wrap-up

**Portfolio Documentation (Week 5 & Week 10):** Combined evidence package including screenshots of completed exercises, research findings, and challenge progress to demonstrate competency development to instructors.

## 5. How are you going to do it (inferencing, delivery)?

### Following IBM Data Science Methodology:

**Phase 1: Proposal (Business Understanding, Analytic Approach, Data Requirements)**

- **Business Understanding**: Entertainment industry investment risk - 60% of films fail to break even

- **Analytic Approach**: Classification model (Hit/Break-even/Flop) using industry-standard profitability thresholds

- **Data Requirements**: Pre-release movie features with historical box office outcomes

**Phase 2: Provisioning (Data Collection, Data Understanding, Data Preparation)**

- **Data Collection**: TMDB API, OMDB API, cleaned up datasets based on algorithms

- **Data Understanding**: Explore movie patterns, assess data quality, identify relationships

- **Data Preparation**: Clean data, handle missing values, create basic features

**Phase 3: Predictions (Modeling, Evaluation)**

- **Modeling**: Start with Nearest Neighbors only (as required for Iteration Zero)

- **Evaluation**: Industry-standard profitability assessment, no advanced validation yet

### Iteration Zero Constraints (Quick & Dirty Proof of Concept):

* **TMDB/OMDB API Integration:** Real-time data collection with proper rate limiting and error handling/ real users rating
* **Nearest Neighbors Only**: No other algorithms, hyperparameter tuning, or cross-validation yet
* **Basic Features**: Simple preprocessing of API responses, no advanced feature engineering
* **Target Variable**: Real-world profitability classification using industry standards

### Competency Development Through Methodology:

**Professional Standard**: Following IBM methodology systematically, clear documentation at each stage

**Personal Leadership**: Independent decisions on data interpretation and model evaluation approach

**Explainable AI**: Nearest Neighbors naturally interpretable - can show which similar movies influenced each prediction

**Data Preparation & Analysis**: Thorough data understanding and systematic preparation following methodology stages

**Model Engineering**: Proper modeling stage implementation with evaluation metrics

### Technical Implications

Domain research supports classification approach over regression for several business-aligned reasons:

- Industry stakeholders require categorical investment decisions (invest/pass) rather than precise revenue predictions

- Classification models provide more interpretable results for business communication - Uncertainty inherent in entertainment industry makes exact revenue prediction unreliable

- Three-category system (Hit/Break-even/Flop) matches actual industry decision-making frameworks

### Iterative Process:

After Iteration Zero proves feasibility, return to Domain Understanding for deeper analysis, then progress through subsequent iterations with more sophisticated techniques, multiple algorithms, and advanced feature engineering.

## Why This Project is Ideal for Competency Development

**Explainable AI Focus:** Movies are perfect for explanation - everyone understands concepts like genres, budgets, and star power, making it easier to explain why certain combinations lead to success or failure.

**Professional Standard Opportunities:** The entertainment industry context provides rich opportunities for professional communication, from technical documentation to business recommendations.

**Personal Leadership Development:** Working in a domain I’m passionate about encourages independent exploration and creative problem-solving approaches.

**Data & Model Engineering:** Movie data offers complex, real-world challenges with categorical variables, missing data, and feature interaction opportunities.

This project represents the perfect intersection of personal passion and competency development, ensuring both engagement and systematic skill demonstration across all required areas.

## References

### Literature Study Sources

[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.sciencedirect.com/science/article/abs/pii/S0957417405001399) *Expert Systems with Applications*, 30(2), 243-254.

### Industry Analysis Sources

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

[The Numbers. (2024). Movie Financial Database.](https://www.the-numbers.com/)

### Competitive Analysis Sources

[Kaggle. (2024). TMDB Box Office Prediction Competition.](https://www.kaggle.com/c/tmdb-box-office-prediction)

### Industry Success Metrics Sources

[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)

### Statistical Data Sources

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

[Statista. (2024). Global Entertainment and Media Market.](https://www.statista.com/outlook/dmo/media/movies/worldwide)

[The Movie Database (TMDB) API Documentation](https://www.themoviedb.org/documentation/api)

[TMDB API v3 Reference](https://developers.themoviedb.org/3/getting-started/introduction)