MOVIE SUCCESS PREDICTION USING SUPERVISED MACHINE LEARNING

SUMIT LAGALE — Bachelor of Computer Applications (BCA)

KLE BK BCA COLLEGE, CHIKODI | Academic Year: 2025–2026

Declaration

I declare that this project, **“Movie Success Prediction Using Supervised Machine Learning”**, submitted to **KLE BK BCA COLLEGE, CHIKODI** for the partial fulfilment of the requirements for the Bachelor of Computer Applications (BCA) degree, is my original work. All sources used have been properly acknowledged and cited in accordance with academic standards. This work has not been submitted previously for any other degree or diploma.

Sincerely,

SUMIT LAGALE

Acknowledgement

I am deeply grateful to my project guide and faculty for their continuous guidance and feedback. My thanks also go to the college administration and department staff for resources and support, and to my peers for their assistance with data collection.

The encouragement from my family and mentors was essential, greatly contributing to the analytical rigor and academic quality of this project.

Table of Contents

Page numbers to be added in final pagination. This TOC is intentionally concise to preserve a clean modern aesthetic.

Abstract

Predicting Movie Success

This study explores supervised learning methods to forecast movie success, building a robust predictive analytics framework for informed decision-making in film production and distribution.

Methodology & Features

We evaluate classification algorithms (Logistic Regression, Decision Tree, Random Forest, SVM, Gradient Boosting) using feature-engineered inputs such as budget, marketing, cast popularity, director rating, genre, and release season. Feature engineering includes encoding, interaction terms, and normalization.

Model Evaluation & Insights

Model performance is assessed using accuracy, precision, recall, F1-score, and AUC-ROC. Comparative analysis highlights the effectiveness of ensemble techniques in reducing commercial risk and guiding resource allocation.

Operational Deployment

Results are discussed in the context of decision support systems for stakeholders, outlining pathways for operational deployment and continuous model refinement through real-time data integration.

This abstract synthesises methodological choices and expected contributions to applied business intelligence within the film industry.

Introduction

The film industry is characterised by high capital intensity and substantial commercial risk: a small fraction of releases recoup production and marketing investments. Predictive analytics provides the capacity to quantify risk, improve investment decisions and optimise marketing allocation. Machine learning offers scalable tools for extracting patterns from heterogeneous data—financial, demographic and social signals—so producers and distributors can make evidence-based decisions.

This project situates supervised machine learning within a business intelligence framework, emphasising classification models that predict binary or categorical measures of success. By converting creative and economic signals into actionable predictors, the system aims to enhance strategic planning and reduce uncertainty in film financing.

Problem Statement & Objectives

Problem Statement

How can supervised machine learning models be developed and validated to reliably predict movie success using production, cast, marketing and temporal features, thereby informing production investment and distribution strategies?

Objectives

1. Collect and curate a representative dataset of feature variables for released films.
2. Engineer predictive features and encode categorical attributes effectively.
3. Develop and train multiple classification algorithms and ensembles.
4. Evaluate models using accuracy, precision, recall, F1-score and AUC-ROC.
5. Compare models to select an optimal decision-support solution for stakeholders.
6. Illustrate deployment pathways and propose enhancements for production use.

Literature Review

Prior studies demonstrate that budget, marketing spend and cast popularity are strong predictors of box-office performance. Research indicates ensemble methods—Random Forests and Gradient Boosting—often surpass single learners by reducing variance and capturing complex non-linear relationships.

Empirical work has used feature selection and dimensionality reduction to mitigate multicollinearity from correlated financial variables. Several investigations integrate social metrics and pre-release sentiment to enhance early prediction accuracy.

Budget & Marketing

Shown repeatedly to have strong effect sizes in regression and classification studies.

Cast Popularity

Metrics derived from prior credits and social following improve model discrimination.

Ensembles

Combine multiple algorithms to deliver robust, generalisable predictions.

Methodology & System Architecture

Data Collection

Gathering data from industry datasets and curated sources to form a comprehensive foundation.

Data Preprocessing

Involves encoding categorical fields (one-hot and ordinal), stratified train-test splitting, and feature scaling (standardisation).

Model Development

Building models across Logistic Regression, Decision Tree, Random Forest, SVM, and Gradient Boosting, with hyperparameter tuning.

Model Evaluation

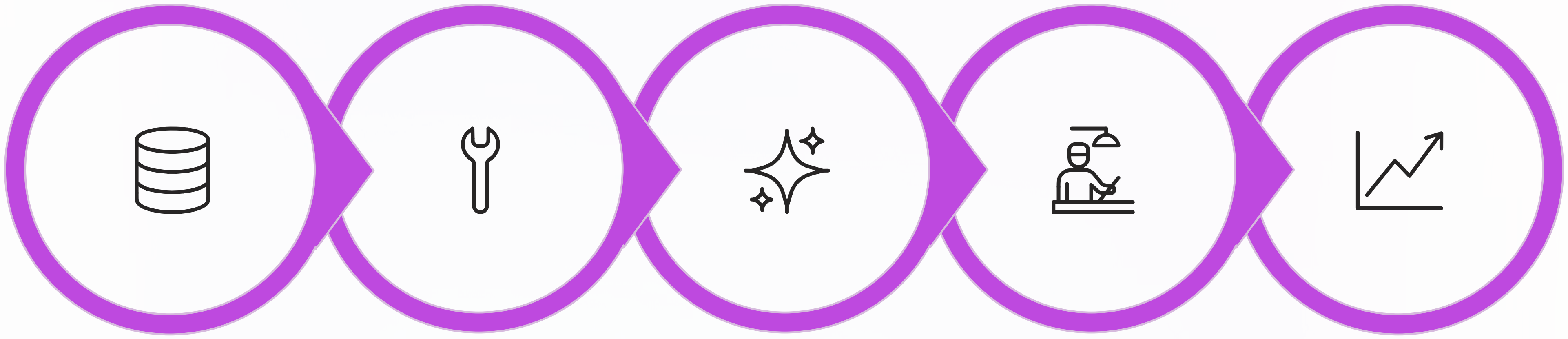
Assessing model performance using cross-validation and key metrics: accuracy, precision, recall, F1, and AUC-ROC.

Model Comparison & Selection

Choosing the optimal model based on generalisation capability, interpretability, and operational constraints.

Deployment Preparation

Finalising the selected model for integration as a robust decision support system.



Data Collection

Preprocessing

Feature Engineering

Model Training

Evaluation

Leave the diagram area as a clear, central architecture placeholder to insert a detailed architecture diagram in the final report.