Concept Note: Movie Revenue Prediction Using Machine Learning

Title

Predicting Movie Revenue Using Machine Learning: A Case Study with the TMDB 5000 Movie Dataset

Problem Statement

The film industry invests millions of dollars into producing and marketing movies. However, predicting whether a movie will succeed at the box office remains uncertain due to the multitude of factors influencing revenue, such as genre, budget, release date, cast, and runtime.

Studios, investors, and distributors seek data-driven insights to assess potential returns before a movie's release. This project aims to build a predictive machine learning model that estimates the box office revenue of a movie using pre-release data. By analyzing and modeling these relationships, the project will help stakeholders make more informed decisions, reduce financial risk, and optimize investments.

Objective

To develop and evaluate machine learning models capable of predicting movie revenue using structured metadata from the TMDB 5000 Movie Dataset.

Dataset

Source: TMDB 5000 Movie Dataset (Kaggle)
Link: `https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata`

The dataset includes metadata for over 5,000 movies such as:

- Budget
- Genre
- Release Date
- Cast and Crew
- Production Companies
- Keywords
- Runtime
- Popularity
- Vote Average and Count

Proposed Models

We will implement and compare multiple regression models to identify the best performer:

- Linear Regression
- Ridge and Lasso Regression
- Random Forest Regressor
- XGBoost Regressor
- Gradient Boosting Regressor

Model selection and tuning will be done using **GridSearchCV** with 5-fold cross-validation to optimize hyperparameters and assess performance.

Evaluation Metrics

To compare model performance, we will use:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R^2 Score

Expected Output

Expected Output

- A clean, preprocessed dataset ready for machine learning.
- A trained model that accurately predicts revenue.
- Feature importance insights to identify the most influential factors.
- A Streamlit web application that allows users to input movie features and receive revenue predictions, along with model insights.