Box Office Revenue Prediction Using Machine Learning: A Comparative Study of Regression Models on the TMDB Dataset

Problem Statement

The film industry invests heavily in movie production and marketing. However, forecasting box office performance remains a complex challenge due to numerous influencing factors, including genre, budget, cast, release timing, and production scale.

This project aims to develop a machine learning framework that accurately predicts movie revenue using metadata available prior to release. By treating revenue prediction as a regression problem, we enable decision-makers—such as investors, studios, and distributors—to make informed funding and marketing decisions.

The project also emphasizes robust feature engineering, interpretability (via SHAP values), and safeguards against temporal leakage. Our models and workflows are designed with production-readiness in mind, paving the way for real-world integration.

Objective

To develop and compare the performance of multiple regression models for predicting box office revenue using metadata from the TMDB dataset, while ensuring interpretability, temporal robustness, and real-world deployment readiness.

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

Feature Engineering and Preprocessing

Key features such as budget, runtime, release date, popularity, cast size, and production companies will be extracted and cleaned.

Preprocessing steps include:

- Handling missing values and outliers
- Encoding categorical variables (e.g., genres, production companies)
- Parsing and extracting date-related features (e.g., month, season)
- Log-transforming skewed variables like budget and revenue
- Ensuring no data leakage from post-release attributes

We will also incorporate feature importance analysis using SHAP to understand the impact of different predictors on the revenue forecast.

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
- SHAP-based interpretability for model explainability

Expected Output

- A clean, production-ready dataset with selected features
- A comparative analysis of regression models with performance scores
- Interpretability insights using SHAP values
- A **Streamlit web application** that allows users to input movie features and receive revenue predictions, with model interpretation visualizations