Box Office Revenue Prediction Analysis

Introduction: Predicting box office revenue is an essential aspect of film industry analytics. This project aims to develop a machine learning model that predicts the domestic revenue of a movie based on various features such as budget, distributor, MPAA rating, and genre. The methodology involves data preprocessing, exploratory data analysis (EDA), feature selection, and model training.

Skills used: Data Sieving, Regression Analysis, Feature Engineering

Dataset Overview: The dataset used for box office revenue prediction contains various attributes describing the movie, its release details, and financial performance.

Dataset Features

1. **title** - Movie title (not used for prediction)
2. **domestic_revenue** (Target Variable) - Total domestic earnings of the movie
3. **distributor** - Studio or company that distributed the movie
4. **opening_theaters** - Number of theaters during the opening weekend
5. **budget** - Total production budget (removed during processing due to missing values)
6. **MPAA** - Movie rating (G, PG, PG-13, R, NC-17)
7. **genres** - Movie genre(s)
8. **release_days** - Number of days the movie was in theaters

The target variable **domestic_revenue** is what we aim to predict based on other attributes, i.e., we consider **domestic_revenue** as the response variable.

Analysis of Data

1. **Data Distribution**
    - The dataset contains 2,694 movies with various distributors, ratings, and genres.
    - Domestic revenue varies widely, with some movies earning significantly more.
    - MPAA ratings impact revenue, with **PG and R-rated movies** having higher average earnings.
    - Movies with a longer release duration tend to earn more revenue.
2. **Feature Importance** Using feature selection techniques, the most influential factors in revenue prediction were identified:
    - **Opening Theaters** - More opening theaters correlate with higher domestic revenue.
    - **MPAA Rating** - Some ratings (e.g., PG and R) perform better in terms of revenue.
    - **Release Duration** - Longer release duration is associated with higher earnings.
    - **Distributor** - Some distributors (e.g., Disney, Warner Bros.) have a history of high-grossing movies.

- **Genres** - Action and animation movies tend to earn higher revenue.

Machine Learning Model Implementation

1. **Data Preprocessing**
    - **Handling Missing Values** - Budget column was removed due to missing data.
    - **Encoding Categorical Variables** - Label encoding was applied to **distributor** and **MPAA rating**.
    - **Text Feature Processing** - Genres were vectorized using **CountVectorizer**.
    - **Feature Scaling** - Applied **StandardScaler** to normalize numerical variables.
2. **Model Selection and Training**
    - **Algorithm Used** - **XGBoost Regressor**, an advanced gradient-boosting algorithm.
    - **Train-Test Split** - 90% training data, 10% testing data.
    - **Hyperparameter Optimization** - Default XGBoost parameters were used for initial training.
3. **Model Performance** The final model was evaluated using **Mean Absolute Error (MAE):**
    - **Training Error (MAE)**: 0.21045
    - **Validation Error (MAE)**: 0.63582

Key Insights from Model Performance

- **XGBoost** performed well, indicating strong predictive power for revenue estimation.
- Opening theater count and release duration were critical features in revenue prediction.
- Encoding categorical variables (**MPAA, distributor**) improved model accuracy.
- Feature engineering (genre vectorization) provided additional predictive value.
- Further tuning, such as hyperparameter **optimization** and additional revenue-related features (e.g., marketing spend), could enhance model performance.

This analysis provides valuable insights into box office revenue prediction and demonstrates how various factors contribute to a movie's financial success.
