Box Office Revenue Prediction

Observations:

From the dataset and preprocessing steps, we can observe that:

- The dataset consists of **2,694 movies** with features such as **title, distributor, MPAA rating, genres, release days, and domestic revenue**.
- The target variable is **domestic_revenue**, and **world_revenue** and **opening_revenue** were removed from the dataset as they were not needed for this prediction.
- The **budget** column was dropped due to missing values.
- Missing values in **MPAA rating** and **genres** were filled with their **mode (most frequent value)**.
- Categorical variables such as **distributor and MPAA rating** were **label-encoded** to convert them into numerical values.
- The **genres column was vectorized** using CountVectorizer, and genres with more than **95% zero values were removed** to reduce sparsity.
- The dataset was split into **training (90%) and validation (10%) sets**.
- The **features were standardized** using **StandardScaler** to improve model performance.

From the **correlation analysis**, we observe that:

- **Opening theaters and release days** are important factors affecting domestic revenue.
- Highly correlated variables were retained as they contribute significantly to the model's predictive power.

Model Performance: By fitting the **XGBoost Regressor**, we observe that:

- The model was trained on **2,424 samples**, and tested on **270 samples**.
- The **Mean Absolute Error (MAE)** on the training set is **0.21045**, indicating that the model fits the training data well.
- The **MAE on the validation set is 0.63582**, which shows a reasonable generalization ability.
- The small gap between training and validation errors suggests **minimal overfitting**.

Conclusion:

- The **XGBoost Regressor** provides **the best performance** for predicting box office revenue.
- The **low MAE values** indicate that the model is effective at estimating domestic revenue.
- The model can be improved further by **hyperparameter tuning** or adding more relevant features such as marketing spend, actor popularity, and franchise status.
