Movie Rating Prediction

1 Introduction

Movie ratings play a crucial role in the entertainment industry by influencing audience preferences, box office revenue, and recommendation systems. Predicting movie ratings using machine learning helps understand the factors that contribute to a movie's success and enables automated rating estimation for new or upcoming films.

This project focuses on building a regression-based machine learning model to predict movie ratings using features such as genre, director, and cast information. By analyzing historical movie data, the model aims to learn patterns that correlate movie attributes with audience or critic ratings.

2 Dataset Description

The dataset consists of the following attributes related to movies, such as:

- Movie genre
- Director
- Lead actors
- Additional numerical or categorical features
- Target variable: movie rating

The target variable is continuous, making this a regression problem.

3 Methodology

3.1 Exploratory Data Analysis

Exploratory data analysis is performed to understand feature distributions and relationships with movie ratings:

- Convert the **Rating** column to numeric format, remove invalid entries, and retain ratings between 1 and 10.
- Extract valid four-digit values from the **Year** column and convert them to numeric format.
- Clean the **Duration** column by removing textual units and converting values to minutes.
- Convert the **Votes** column to numeric format by removing comma separators.

- Remove records with missing or inconsistent values to ensure data quality.
- Visualize the distribution of movie **ratings** using histograms to analyze rating spread and skewness.
- Analyze the **votes** distribution before and after applying a logarithmic transformation to reduce skewness and handle large-scale variations.
- Identify the **top 10 movie genres** by frequency to understand genre popularity.
- Examine relationships between **rating** and numerical features such as **year**, **duration**, and **log-transformed votes** using scatter plots.

3.2 Feature Transformation

A preprocessing pipeline is created to prepare features for model training:

- Numerical features (*Year*, *Duration*) are standardized using **StandardScaler**.
- Categorical features (*Genre*, *Director*, *Actor 1*) are encoded using **One-Hot Encoding** with unknown categories handled safely. Then, the dataset is split into training and testing sets to evaluate model performance

3.3 Regression Models Used

The following regression models are trained to predict movie ratings:

- Linear Regression
- Random Forest Regressor
- Gradient Boosting Regressor

3.4 Model Training

Regression techniques are used to predict movie ratings. The model is trained on the processed training dataset to learn the relationship between movie features and their ratings.

3.5 Evaluation Metrics

The model's performance is evaluated using standard regression metrics, including:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2 Score)

These metrics help measure how closely the predicted ratings match the actual ratings.

4 Results

Table 1: Performance Comparison of Regression Models

Model	**MAE**	**MSE**	**RMSE**	$\mathbf{R^2}$
Linear Regression	1.1209	2.1744	1.4746	-0.1696
Random Forest Regressor	0.9041	1.3956	1.1813	0.2494
Gradient Boosting Regressor	0.9428	1.4436	1.2015	0.2235

The experimental results indicate that ensemble-based models outperform linear regression for movie rating prediction. Linear Regression shows the weakest performance with a negative R^2 value, suggesting that it fails to capture the underlying patterns in the data. Random Forest Regressor achieves the lowest error values and the highest R^2 score, making it the best-performing model in this study. Gradient Boosting Regressor also improves upon linear regression but performs slightly worse than Random Forest, possibly due to sensitivity to feature noise and limited data representation. Hence, **Random Forest** is chosem as the best model.

5 Conclusion

This project successfully demonstrates the application of machine learning regression techniques to predict movie ratings using historical data. Through data preprocessing, feature engineering, and model training, a predictive system was developed that estimates movie ratings with acceptable accuracy.

The results highlight the importance of structured movie attributes in determining ratings and showcase the potential of machine learning in recommendation and decision-support systems. Future improvements may include incorporating additional features such as budget, release year, and audience demographics, as well as experimenting with advanced models like ensemble regressors or neural networks.