Predicting Movie Release Year and Genres

Kashyap Raparthi

May 2, 2024

Abstract

This report details the process of developing a machine learning model that predicts the release year and genres of a movie directed by a given director. The project utilized a provided movie metadata dataset. The methodology involved data preprocessing, feature engineering, model development, and evaluation.

1 Introduction

This project aimed to develop a machine learning model capable of predicting the release year and genres of a movie based on the director's information. The project utilized a dataset containing movie metadata.

2 Methodology

The methodology followed these key steps:

2.1 Data Preprocessing

1. Data Loading and Exploration: The movie metadata was loaded using libraries like pandas and analyzed for structure and content.
2. Data Cleaning: Duplicate entries were removed. Irrelevant columns were dropped.
3. Handling Missing Values: Rows with missing values in crucial features were dropped. Missing values in some features were imputed using techniques like finding the most frequent value (mode).

2.2 Feature Engineering

New features were created to capture insights from existing data:

1. director_avg_rating: Average IMDB score of a director's movies.
2. director_movie_count: Number of movies directed by a particular director.

3. years_since_last_movie: Difference between the current year and the director's last movie's release year.

The correlation between different features was analysed in order to see which features varied the title_year, which showed that imdb_score , num_critic_reaviews and years_since_last_movie had a significant correlation with title_year.

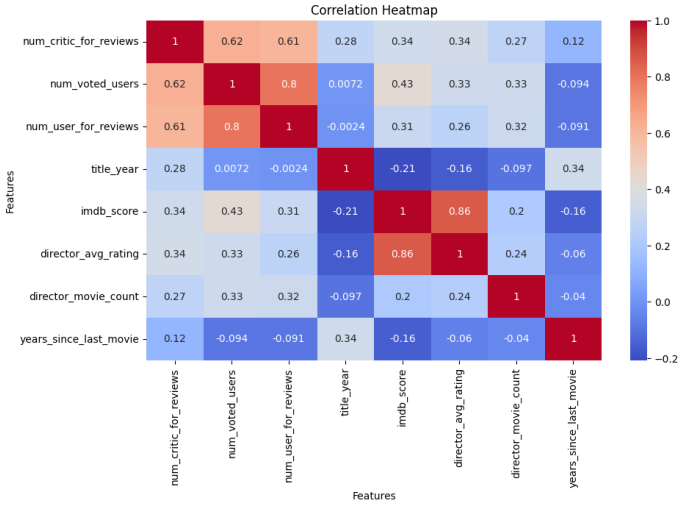

Figure 1: Correlation

2.3 Data Splitting

The preprocessed data was split into features (X) and target variables (y):

1. Release Year: Target variable 'y_release_year' contained the movie release year.
2. Genres: Target variable 'y_genres' was a multi-label categorical variable containing movie genres, one-hot encoded.
3. The data was further split into training and testing sets for model training and evaluation.

2.4 Preprocessing Pipeline

Separate pipelines were created for numerical and categorical features:

1. Numerical features were handled using imputation and standardization.
2. Categorical features were handled using imputation and one-hot encoding.
3. A 'ColumnTransformer' combined these pipelines for consistent preprocessing.

2.5 Model Development

Two separate models were developed:

1. Release Year Prediction: A Gradient Boosting Regressor was chosen for its ability to handle non-linear relationships.
2. Genres Prediction: A MultiOutputRegressor with an underlying XGBoost Classifier was chosen for multi-label classification.

2.6 Model Training

The preprocessed training data was used to train both models.

2.7 Model Evaluation & Results

The models were evaluated on unseen testing data:

1. Release Year Prediction: Mean Absolute Error (MAE) measured the average difference between predicted and actual release years.
2. Genres Prediction: F1 score (micro-averaged) evaluated the overall multi-label classification accuracy.

The MAE was about 4 years and F1 score about 0.47.

3 Conclusion

This project successfully developed and evaluated a machine learning model for predicting movie release year and genres based on director information. The methodology employed data cleaning, feature engineering, preprocessing pipelines, and hyperparameter-tuned models. Future work could involve incorporating additional features, exploring different model architectures, and potentially using ensemble learning for further improvement.