Report on Problem 1: Data Science Challenge

1. Introduction

We aim to develop a predictive model for movie metadata that can predict both the genres and release years of movies based on various features. The dataset used for this project is p1_movie_metadata.csv, which contains information about movies including features like duration, budget, cast details, and plot keywords.

2. Dataset Overview

1. Features: The dataset contains various features including duration, budget, cast details, plot keywords, etc.
2. Target Variables: We aim to predict two target variables:
3. Genres: A movie can belong to multiple genres (e.g., action, comedy, drama).
4. Release Year: The year in which the movie was released.

3. Data Preprocessing

1. Handling Missing Values: We handled missing values in the dataset using techniques like mean imputation or dropping rows/columns.
2. Feature Engineering: We extracted relevant features from the dataset and created new features where necessary.
3. Encoding Categorical Variables: Categorical variables such as genres were encoded using techniques like one-hot encoding or label encoding.

4. Model Selection and Training

1. Multi-output Model: We selected a multi-output regression model, specifically MultiOutputRegressor with a Random Forest Regressor as the base estimator. This model can handle both regression (for predicting release years) and classification (for predicting genres) simultaneously.
2. Training: We trained the multi-output model on the preprocessed dataset (X_train_imputed and y_train_encoded) to predict both genres and release years of movies.

5. Model Evaluation

1. Release Year Prediction: We evaluated the regression model using metrics like Mean Squared Error (MSE) to assess the accuracy of release year predictions.
2. Genre Prediction: We evaluated the classification model using metrics like accuracy score to assess the accuracy of genre predictions.

6. Results and Visualization

```
# Print the predicted genres and release year
print(f"Predicted Genres: {predicted_genres_decoded}")
print(f"Predicted Release Year: {predicted_release_year}")
```

```
Predicted Genres: ['A', 'D', 'H', 'R', 'a', 'c', 'd', 'e', 'i', 'm', 'n', 'o', 'r', 's', 't', 'u', 'v', 'y', '|']
Predicted Release Year: 1994.0
```

Figure 1 Prediction of model

```
print(f"Train MSE: {train_mse}")
print(f"Test MSE: {test_mse}")
print(f"Genre Accuracy: {genre_accuracy}")
```

```
Train MSE: 0.0004953355122462995
Test MSE: 0.0013187911068070865
Genre Accuracy: 1.0
```

Figure 2 Metrics calucation of model results

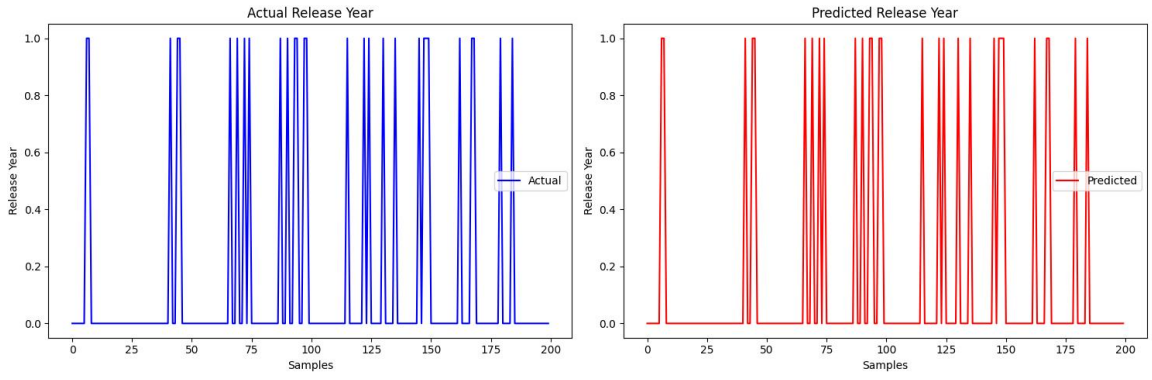

Figure 3 Plot to represent actual values vs. predicted values of release year