Data Science Challenge

Objective:

The problem includes prediction of genres as well as release year of a movie given its director's name.

Methodology:

The process can be summarized in these steps:

1) Data Loading and Cleaning:
   a) The data is first loaded from a CSV file.
   b) Any rows containing missing values (NaN) are removed to ensure clean data for training.
2) Pre-processing for Prediction Tasks:
   a) The 'director_name' column, a string data type, serves as the main input for both prediction tasks (year and genre). To prepare it for modeling, unique director names are assigned integer labels.
   b) Year Prediction:
      i) Numerical columns are analyzed to find correlations relevant to predicting the release year.
      ii) Columns with a correlation coefficient above 0.1 with the 'title_year' column are considered relevant features.
      iii) These chosen features are combined to create a feature vector used for training the year prediction model.
      iv) Since only the director's name is available during testing, additional information is needed at test time. The model addresses this by finding the average values of relevant features at training time for each director. When a new director name is provided for testing, a feature vector containing these average values is constructed for prediction.
   c) Genre Prediction:
      i) The 'genres' column contains textual data with various genres, making correlation analysis unsuitable.
      ii) Therefore, only the 'director_name' column is used as input for the genre prediction model.
      iii) Each genre within the 'genres' column value is identified, and separate entries are created for each movie-genre combination. This expands the dataset size.
      iv) Genres are assigned unique integer labels using a dictionary for efficient processing.
3) Training and Testing:
   a) After pre-processing, the data is split into training and testing sets, typically with an 80/20 split (80% for training and 20% for testing).
   b) During testing, only the 'director_name' is provided as input to the model. The models are designed to handle any necessary feature engineering internally and predict the corresponding output (year or genre).
4) Model for predicting year of release:
   a) The model used to predict movie release years based on director names and other relevant data is a Random Forest Regressor, an ensemble learning technique well-suited for regression tasks.

b) Random Forest is an ensemble method that combines predictions from multiple decision trees. Its key aspects are as follows:
   i) *Decision Trees:* These are individual tree-like models that learn to make predictions by splitting the data based on a series of conditions.
   ii) *Ensemble Learning:* The Random Forest combines predictions from a large number (e.g., 300 in this case) of these decision trees. Each tree is trained on a random subset of the data and uses a random selection of features at each split point. This helps to reduce variance and improve the overall accuracy of the model.

c) This approach utilizes a Random Forest model trained with director names and relevant features to predict movie release years. The model can handle missing features during testing by employing pre-computed average values for each director.

5) Model for predicting genre:
   a) This document details the approach used to predict movie genres based on director names using the K-Nearest Neighbors (KNN) algorithm. KNN is a non-parametric machine learning technique widely used for classification tasks.
   b) When predicting the genre for a new movie, the KNN model calculates the distance between the director name (represented numerically) of the new movie and all director names in the training data.Various distance metrics can be used. We have used Euclidean distance being a common choice.
   c) A predefined value of K is chosen (K = 12 in this case). This value represents the number of closest neighbors to consider for prediction. After calculating distances to all training data points, the K data points with the smallest distances (nearest neighbors) to the new movie's director name are identified.
   d) Each of the K nearest neighbors likely has associated genres from the movies they directed in the training data. The genres of these K nearest neighbors are analyzed. The genre that appears most frequently among these neighbors is considered the predicted genre for the new movie. If the K nearest neighbors directed movies with the genres "Comedy" (3 times), "Drama" (2 times), and "Action" (1 time), then "Comedy" would be the predicted genre for the new movie.

Result:

Various hyper parameter tuning such as changing value of number of decision trees and number of nearest neighbors is done. The best results are shown below:

For predicting year of release:

Loss = 84.447

For predicting genre:

Accuracy = 19.61(%)

Conclusion:

The results indicate that the models performed better at predicting genre than year of release. The model seems to have a significant average error in predicting release years. The model for predicting genre is not ideal but it suggests some ability to classify genres based on director names. The performance of task can be further improved by including more features and using other models such as Support Vector Machines (SVM).