**Phase 5: Project documentation and submission**

**Project title: Predicting IMDb Scores**

**Dataset Link:** https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores

**Problem Statement**

The problem is to develop a machine learning model that predicts IMDb scores of movies available on Films based on features like genre, premiere date, runtime, and language. The objective is to create a model that accurately estimates the popularity of movies, helping users discover highly rated films that match their preferences. This project involves data preprocessing, feature engineering, model selection, training, and evaluation.

**Design Thinking**

**1.Data Source**

The data source for this project will be the dataset available at the provided link: [Netflix Original Films IMDb Scores Dataset](https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores). This dataset contains essential information about movies, including IMDb scores, genre, premiere date, runtime, language, and more. It will serve as the foundation for building the predictive model.

**2.Data Preprocessing**

Data Preprocessing is a vital step in ensuring the quality and reliability of the dataset. The following actions will be taken during this phase:

* Data Cleaning: Identify and address missing or inconsistent data to ensure a clean dataset.
* Data Transformation: Convert categorical features such as genre and language into numerical representations using techniques like one-hot encoding or label encoding.
* Feature Scaling: Normalize or standardize numerical features to bring them to a consistent scale.

**3.Feature Engineering**

Feature Engineering aims to enhance the dataset's quality and enable the model to make more accurate predictions. Some key feature engineering strategies include:

* Extracting relevant information from the premiere date, such as year or month, to capture any time-related patterns that may affect IMDb scores.
* Creating new features based on genre information, such as the count of genres a movie belongs to.
* Handling outliers, if present, through transformations or removal to reduce their impact on the model.

**4.Model Selection**

Selecting an appropriate regression model is a pivotal decision. Several regression algorithms will be considered, including:

* Linear Regression: A simple and interpretable model to establish a baseline.
* Random Forest Regressor: A versatile ensemble model that can capture non-linear relationships.
* Support Vector Regression: Effective for handling complex datasets and outliers.
* Gradient Boosting Regressor: A powerful ensemble method for predictive accuracy.

The choice of the model will depend on factors like dataset size, complexity, and the model's performance during preliminary evaluations.

**5.Model Training**

Once the regression model is selected, it will be trained using the preprocessed data. The dataset will be divided into a training set and a testing set to assess the model's performance. Hyperparameter tuning may be necessary to optimize the model's predictive capabilities.

**6.Model Evaluation**

Model Evaluation is essential to determine how well the model predicts IMDb scores. Key evaluation metrics will include:

* Mean Absolute Error (MAE): This measures the average absolute difference between predicted and actual IMDb scores.
* Mean Squared Error (MSE): It quantifies the average squared difference between predicted and actual IMDb scores.
* R-squared (R2): An indicator of how well the model explains the variance in IMDb scores.

These metrics will provide insights into the model's accuracy and its ability to meet the project's objectives.

**Why Linear regression in Model Training?**

Linear Regression can be a valid choice for a baseline model when predicting IMDb scores, especially if you want interpretability and computational efficiency. However, it's essential to assess the model's performance and consider more complex models if the linear assumption does not hold or if you aim for higher predictive accuracy. The choice of model should be based on the specific characteristics of the dataset and the project's objectives.

* **Interpretability:** Linear Regression is a straightforward and interpretable model. It provides clear insights into the relationships between input features (such as genre, premiere date, runtime, and language) and the IMDb scores. This interpretability can be valuable when you want to understand the impact of individual features on the outcome.
* **Simplicity:** Linear Regression assumes a linear relationship between the input features and the target variable (IMDb scores). While this is a simplification, it can be effective when the relationship is primarily linear or close to linear. Linear models are less prone to overfitting in such cases, making them a good starting point.
* **Speed and Efficiency:** Linear Regression is computationally efficient and can be trained relatively quickly, especially on large datasets. This efficiency can be advantageous when you want to perform quick initial experiments or require real-time predictions.
* **Baseline Model:** Linear Regression can serve as a useful baseline model. You can start with a Linear Regression model, assess its performance, and then compare it to more complex models like Random Forest or Gradient Boosting. If Linear Regression provides satisfactory results, it may save computational resources and suffice for your prediction task.

**Conclusion**

This document outlines the initial steps for addressing the problem of predicting IMDb scores for Netflix Original Films. By utilizing the provided dataset and following the data preprocessing, feature engineering, model selection, training, and evaluation phases, the project aims to deliver a predictive model that enhances the user experience by recommending highly rated films. Subsequent phases of the project will involve the practical implementation of these steps, fine-tuning the model, and providing a valuable solution for Netflix Original Films enthusiasts.