PHASE 1: PROBLEM DEFINITION AND DESIGN THINKING

In this part you will need to understand the problem statement and create a document on what have you understood and how will you proceed ahead with solving the problem. Please think on a design and present in form of a document.

PROBLEM DEFINITION:

The "Predicting IMDb Movie Scores" project seeks to harness the power of machine learning to forecast IMDb ratings for movies. IMDb scores play a pivotal role in guiding viewers' movie choices and are of paramount importance to the film industry. This project endeavours to construct an accurate predictive model that can provide insights to filmmakers, distributors, and movie enthusiasts. To achieve this, we collect and preprocess extensive movie-related data, explore patterns through data analysis, employ machine learning algorithms, and ultimately deliver a model capable of estimating IMDb scores. The findings and methodologies unveiled in this project hold the potential to enhance the understanding of the factors that influence movie ratings and revolutionize the movie industry's decision-making processes.

The goal of this project is to develop a predictive model that can accurately estimate IMDb scores for movies. IMDb scores are a widely recognized metric for evaluating the quality of movies, and accurate prediction can provide valuable insights to movie studios, distributors, and viewers alike. The project aims to leverage machine learning techniques to build a robust prediction model.

With reference to the link **https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores**

DESIGN THINKING:

1. DATA SOURCE:

Predicting IMDb scores is a complex task that typically involves a variety of features and data sources. Here are some common data sources and features that can be used to predict IMDb scores:

- Movie Metadata
- Cast and Crew Data
- User Reviews
- Box Office Data
- Awards and Nominations
- Movie Trailer Views
- Social Media Mentions
- Rotten Tomatoes or Metacritic Scores
- Historical IMDb Data
- Budget and Production Costs

To create a predictive model for IMDb scores, you would typically gather a dataset that includes these features and their corresponding IMDb scores. Machine learning techniques, such as regression or ensemble models, can then be applied to train the model for prediction. Keep in mind that feature engineering and data preprocessing are crucial steps in building an accurate prediction model. Additionally, regularly updated data is important for maintaining the model's accuracy over time.

2. DATA PROCESSING

Data cleaning is a crucial step in the data preprocessing pipeline in data science. Data preprocessing means to prepare the data for classification. Data is processed according to the requirements of classification. It involves identifying and addressing issues in the dataset to ensure that the data is accurate, consistent, and ready for analysis or model training. Here, for preprocessing the data, instances with missing attributes are removed Here's a detailed explanation of the data cleaning process

- Handling Missing Data
- Handling Duplicate Data
- Correcting Inaccurate Data
- Standardizing Data
- Handling Outliers
- Dealing with Inconsistent Data Types
- Addressing Data Integrity Issues
- Handling Text Data
- Documenting Changes
- Data Imputation for Time Series Data
- Data Scaling and Normalization

Data cleaning is an iterative process, and it's essential to thoroughly understand the data and the domain to make informed decisions during each step. Careful data cleaning ensures that the subsequent analysis or modelling steps are based on high-quality data, leading to more reliable and meaningful results.

3. FEATURE ENGINEERING

Feature engineering is critical when predicting IMDb scores for movies. Here are some specific feature engineering techniques you can apply:

- Genre Features
- Director and Actor Influence
- Release Date Features
- Budget and Box Office
- Runtime
- Word Analysis from Plot and Reviews
- Awards and Nominations
- User Ratings from Other Platforms
- Movie Sequels and Franchises
- Cultural and Historical Events
- Studio Influence
- Directorial Debut
- MPAA Rating
- Movie Budget to Revenue Ration

Remember to perform thorough data analysis and feature selection to ensure that the engineered features are relevant and do not introduce noise into the model. Experiment with different combinations of features and machine learning algorithms to find the best approach for predicting IMDb scores accurately.

4. MODEL SELECTION

Selecting the right model for predicting IMDb scores or analyzing IMDb data involves understanding your specific problem, the data you have, and your objectives. IMDb scores typically involve regression tasks, where you predict a continuous target variable (movie ratings). Here's a model selection process tailored to IMDb score prediction:

Consider a variety of regression models suitable for IMDb score prediction:

- Linear Regression: A simple model assuming a linear relationship between features and IMDb scores.
- Tree-Based Models: Decision Trees, Random Forest, or Gradient Boosting models can capture non-linear relationships.
- Neural Networks: Deep learning models, like feedforward neural networks or recurrent neural networks (RNNs), can handle complex patterns.

- Support Vector Machines (SVM): SVMs can work well for regression tasks when properly tuned.
- K-Nearest Neighbors (KNN): Suitable for locally patterned data.

5. MODEL TRAINING

Model training for predicting IMDb scores typically involves a regression task, as you're trying to predict a continuous target variable (movie ratings).

- Use the training data to train your selected regression model. The model will learn to make predictions based on the features and target IMDb scores.
- Tune the hyperparameters of your model using techniques like grid search or random search to optimize its performance
- Assess your model's performance on the validation set using the chosen evaluation metric(s). This helps you fine-tune hyperparameters and check for overfitting.
- After tuning your model and validating its performance, evaluate it on the test set to assess how well it generalizes to unseen data. Use the same evaluation metrics as in the validation phase.

6. EVALUATION

Evaluating the performance of a predictive model for IMDb scores is crucial to understand how well the model is performing and to make any necessary improvements. Common evaluation metrics for regression tasks like predicting IMDb scores include:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE
- Root Mean Squared Error (RMSE
- R-squared (R^2) or Coefficient of Determination
- Adjusted R-squared
- Mean Absolute Percentage Error (MAPE

It's essential to choose evaluation metrics that align with your specific goals and the nature of your IMDb score prediction task. For example, if your goal is to recommend movies with high IMDb scores to users, you might prioritize metrics that emphasize accuracy, such as MAE or RMSE. Additionally, it's a good practice to compare your model's performance to a baseline model or a simple heuristic to gauge its effectiveness.

Remember that no single metric provides a complete picture of model performance, so it's valuable to examine multiple metrics and consider the practical implications of your model's performance in real-world scenarios.