**Project Documentation:**

**IMDb Score Prediction**

**Problem Statement:**

The IMDb Score Prediction project aims to develop a model that predicts the IMDb score of movies based on various features. The problem can be framed as a regression task, where the goal is to create an accurate model for predicting IMDb scores for movies and to check the accuracy of the model by algorithms.

**Design Thinking Process :**

**Understanding the Problem:**

We began by understanding the importance of predicting IMDb scores for movies and how it can benefit the film industry and viewers.

**Data Collection:**

We gathered data from various sources, including IMDb, and selected relevant features for prediction.

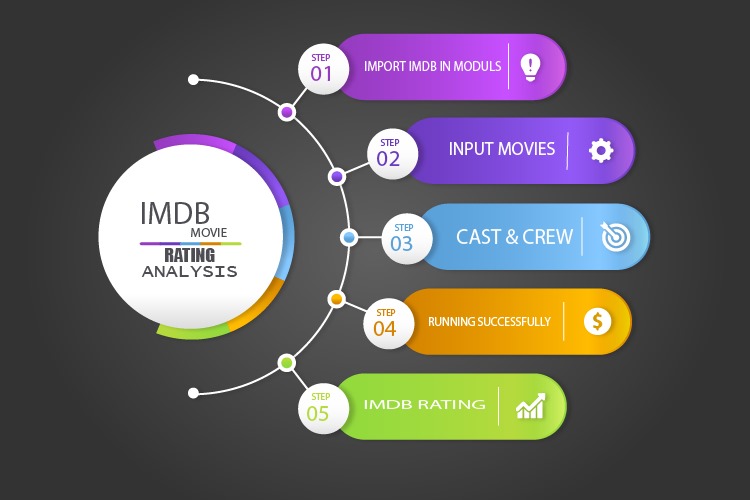
**Data Preprocessing:**

We cleaned the data, handled missing values, and encoded(converted language and movies time as ratings from 1 to 5) categorical variables.

**Model Development:**

We experimented with different regression algorithms to find the most suitable one for this task.

**Model Evaluation:** We used appropriate evaluation metrics to assess the model's performance.



**Phases of Development**

**1.Data Collection and Exploration:**

Acquiring data, understanding its structure, and performing exploratory data analysis.

Data set :[**https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores**](https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores)

**Source code :** <https://github.com/ArBhagavath02/IBM-Naan-Mudhalvan-/blob/main/IMDB%20predicting%20score%20analysis.ipynb>

**2. Data Pre-processing:**

Cleaning, imputing missing values(some values taken down to mean and some data is removed), and transforming data to make it suitable for modelling(making most about the numerical values).

**Source code : https://github.com/ArBhagavath02/IBM-Naan-Mudhalvan-/blob/main/ADS\_Phase2.docx**

**3.Model Selection:** Trying various regression algorithms (e.g., Linear Regression, Random Forest, Gradient Boosting. R-squared algorithm) and selecting the most suitable one.

**Source code : https://github.com/ArBhagavath02/IBM-Naan-Mudhalvan-/blob/main/ADS\_Phase3.ipynb**

**4.Model Training:**

Training the selected model using a training dataset by dividing the data into training and testing data (approx. 70% training and 30% testing).

**Source code : https://github.com/ArBhagavath02/IBM-Naan-Mudhalvan-/blob/main/ADS\_Phase4-1.ipynb**

**5.Model Evaluation:**

Assessing the model's performance using evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

**Dataset Description**

The dataset used for this project contains information about movies, including features like Title, Genre, premiere, Runtime of movie, language. It includes IMDb scores as the target variable.

**Data Pre-processing Steps**

**1.** Handling missing data by imputing or removing rows.

**2.** Encoding categorical features using techniques like one-hot encoding.

**3.** Scaling and normalizing numerical features.

**Model Training Process**

We used a linear regression to train the model on the pre-processed dataset. We split the data into training and testing sets and used cross-validation to tune hyperparameters for optimal performance.

**Choice of Regression Algorithm and Evaluation Metrics**

We selected Linear algorithm for regression due to its simplicity and good performance. For evaluation, we used metrics like R-squared to assess the model's accuracy and generalization.