PROJECT TITLE

PREDICTING IMDB SCORES

Register number: 713921104302

Naan Mudhalvan ID: au71391104302

Email ID: harivignesh545@gmail.com

Name: Harivignesh P

PHASE 1: Problem definition and Design thinking

Problem Definition:

Predicting IMDB scores for movies using applied data science techniques involves various steps. It typically includes data collection, cleaning, feature engineering, and model training. In this case, for the movie "Naan Muthalvan," you'd need a dataset with relevant features like director, actors, genre, release date.

Once you have the data, you can apply regression or classification models to predict the IMDb score. Common techniques include linear regression, decision trees, random forests, or even more advanced methods like neural networks.

 Keep in mind that the accuracy of your prediction will largely depend on the quality and quantity of your data, as well as the effectiveness of the features you choose. Additionally, consider using techniques like cross-validation to evaluate the performance of your models.

Design Thinking:

**Data Source: Netflix Original Films & IMDB Scores**

The initial phase of our IMDB Scores project involves comprehensive data collection to build a robust foundation for our predictive model. IMDb itself provides datasets for public use through their IMDb Datasets page. These datasets include a wide array of information about movies, including ratings.

UCI Machine Learning Repository:

 The UCI Machine Learning Repository is a collection of databases, domain theories, and datasets often used by researchers in machine learning. They may have datasets related to movies and ratings.

MovieLens:

MovieLens is a movie recommendation service that also provides datasets for research purposes. These datasets include movie ratings and other relevant information.

Open Data Platforms:

Some governments or organizations provide open data platforms where you might find datasets related to movies and entertainment, including IMDb scores.

Web Scraping:

If you have some programming skills, you can scrape IMDb or similar movie databases directly using web scraping libraries like BeautifulSoup

Remember to check the licensing and terms of use for any dataset you download to ensure it's suitable for your specific use case. Always give proper attribution if required

**Data Preprocessing: Cleaning and Transformation**

Here are some common tasks a data preprocessor might perform for predicting IMDb scores:

Handling Missing Values:

Identifying and filling in or imputing missing values in the dataset. This can be critical to prevent errors during model training.

Removing Duplicates:

Identifying and removing duplicate records, if any, to ensure that each data point is unique and contributes meaningfully to the model.

Encoding Categorical Variables:

Converting categorical variables (like genres, directors, actors) into a numerical format that can be used by machine learning algorithms. This may involve techniques like one-hot encoding or label encoding.

Scaling and Normalization:

Rescaling numerical features to a similar scale, which can improve the performance of some machine learning algorithms. Techniques like Min-Max scaling or Z-score normalization may be used.

**Feature engineering: Creation of additional features**

Creating new features or transforming existing ones to extract more information that might be relevant for predicting IMDb scores. For example, creating a feature for the number of famous actors in a movie.

Handling Outliers:

Detecting and dealing with outliers that might skew the predictions. This can be important to ensure the model is robust to extreme values.

Date-Time Handling:

Extracting relevant information from date-time variables, like year, month, day, etc., which can be used as features in the model.

**Model Selection: Choosing Regression algorithm**

When selecting a model for predicting IMDb scores in applied data science, you have several options to consider. Here are some popular choices:

Linear Regression:

Pros: Simple, interpretable, and easy to implement. Can provide a good baseline.

Cons: Assumes linear relationships, which may not hold in complex scenarios.

Random Forest Regression:

Pros: Handles non-linear relationships well, provides feature importance, robust to outliers and missing data.

Cons: Can be computationally intensive and might require tuning.

**Model Training: Training the model**

Use the training data to train the chosen model. The model learns to make predictions based on the features provided.

Model Evaluation:

Use the testing set to evaluate the model's performance. Common metrics for regression tasks include Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

Hyperparameter Tuning:

Fine-tune the model's hyperparameters to improve its performance. This may involve techniques like cross-validation.

Model Deployment:

Once satisfied with the model's performance, you can deploy it for making predictions on new data.

Monitoring and Maintenance:

Continuously monitor the model's performance over time. If necessary, retrain the model with updated data.

Remember, the effectiveness of your model will heavily depend on the quality of the data and the features you choose. Additionally, it's important to consider ethical implications, especially when using predictive models in areas like entertainment where subjective biases can come into play.

**Evaluation: Evaluating models performance**

The evaluation of a model for predicting IMDb scores in data science involves assessing its performance and accuracy in estimating movie ratings. Here are some common evaluation metrics and methods:

Mean Absolute Error (MAE): This measures the average absolute difference between the predicted and actual IMDb scores. Lower MAE values indicate better performance.

Mean Squared Error (MSE): This calculates the average of the squared differences between predicted and actual scores. MSE gives higher penalties for larger errors.

Root Mean Squared Error (RMSE): This is the square root of MSE, which provides an interpretable scale similar to the target variable.

R-squared (R2) Score: This measures the proportion of the variance in the dependent variable (IMDb scores) that is predictable from the independent variables (features). R2 score ranges from 0 to 1, where higher values indicate better fit.

Cross-Validation: This technique involves splitting the dataset into multiple subsets (folds), training the model on some and testing on others. It provides a more robust estimate of model performance