IBM NAANMUDHALVAN

APPLIED DATA SCIENCE - PHASE 1
PREDICTING IMDb SCORES

ABSTRACT:

This research introduces a machine learning model tailored for predicting IMDb scores of movies available on the Films platform. By utilizing attributes like genre, premiere date, runtime, and language, the model aims to deliver precise assessments of a movie's popularity. The objective of this predictive tool is to aid users in discovering top-rated films that resonate with their specific tastes, thereby enhancing their overall viewing satisfaction.

DESIGN THINKING:

(1)Data Source Utilization:

Comprehensive Movie Dataset:

The dataset selected for this study comprises a wide array of information pertaining to movies. Key attributes include genre, premiere date, runtime, language, and IMDb scores.

(2)Data Preprocessing:

Data Cleaning and Imputation:

Rigorous procedures were applied to rectify any discrepancies or anomalies in the dataset. Missing values were addressed through appropriate imputation techniques to ensure data completeness.

Categorical to Numerical Conversion:

Categorical features, such as genre and language, were transformed into numerical representations to facilitate seamless integration into the predictive model.

(3)Feature Engineering:

Temporal Features:

In addition to premiere date, temporal features like month and season were extracted to capture potential time-based trends in IMDb scores.

Genre Encoding:

The genre information was encoded using techniques like one-hot encoding to create a structured representation for each movie.

(4)Model Selection:

Algorithmic Choices:

Linear Regression and Random Forest Regressor were identified as suitable regression algorithms for predicting IMDb scores. Linear Regression provides a baseline while Random Forest Regressor offers more complexity and potential for capturing non-linear relationships.

(5)Model Training:

Utilization of Preprocessed Data:

The data, having undergone meticulous cleaning and feature engineering, was used to train the selected regression models.

(6)Evaluation:

Regression Metrics:

The performance of the models was assessed using widely recognized regression metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. These metrics provide a comprehensive assessment of the predictive accuracy and goodness of fit.

Cross-Validation:

To ensure robustness, the models were evaluated using cross-validation techniques, which involve partitioning the dataset into multiple subsets for training and validation.

Interpretation of Results:

The results obtained from the evaluation process were meticulously interpreted to glean insights into the models' predictive capabilities and areas for potential refinement.This comprehensive methodology was implemented to develop and evaluate a predictive model for estimating IMDb scores based on movie features. It encompasses rigorous datapreprocessing, judicious feature engineering, informed model selection, and thorough evaluation to ensure accurate and reliable predictions.