**Predicting IMDb Scores**

**Project Title:** IMDb Score Prediction

**Problem Statement:** Develop a machine learning model to predict the IMDb scores of movies available on Films based on their genre, premiere date, runtime, and language. The model aims to accurately estimate the popularity of movies to assist users in discovering highly rated films that align with their preferences.

**Project Steps**

**Phase 1: Problem Definition and Design Thinking**

**Problem Definition:** 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**: Utilize a dataset containing information about movies, including features like genre, premiere date, runtime, language, and IMDb scores.
2. **Data Preprocessing**: Clean and preprocess the data, handle missing values, and convert categorical features into numerical representations.
3. **Feature Engineering**: Extract relevant features from the available data that could contribute to predicting IMDb scores.
4. **Model Selection**: Choose appropriate regression algorithms (e.g., Linear Regression, Random Forest Regressor) for predicting IMDb scores.
5. **Model Training**: Train the selected model using the preprocessed data.
6. **Evaluation**: Evaluate the model's performance using regression metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

**Project Description:**

**Objective:** The IMDb Scores Prediction project aims to develop a machine learning model capable of predicting IMDb movie ratings with a high degree of accuracy. This predictive model will be a valuable tool for filmmakers, studios, and movie enthusiasts to anticipate the potential success of a movie.

**Scope:** This project will focus on building a predictive model using historical IMDb data, encompassing various movie attributes such as cast and crew information, genre, budget, and release date. The model will aim to forecast IMDb scores for both upcoming and existing movies.

**Data Sources:** We will gather the necessary data from IMDb's extensive database and other reliable sources, including movie industry datasets and online movie databases.

**Steps:**

Data cleaning is a crucial step in any data analysis or prediction project, including an IMDb prediction project. Clean data is essential for building accurate and reliable machine learning models. Here's a general outline of the data cleaning process for an IMDb prediction project:

1. **Data Collection and Inspection:**
   * Gather the IMDb dataset or data from reliable sources.
   * Inspect the data to understand its structure, features, and potential issues.
2. **Handling Missing Values:**
   * Identify missing values in the dataset.
   * Decide on an appropriate strategy for handling missing data, which may include:
     + Removing rows or columns with too many missing values.
     + Imputing missing values using mean, median, mode, or more advanced methods like regression or decision trees.
3. **Dealing with Duplicate Data:**
   * Check for and remove duplicate entries if they exist.
   * Duplicates can distort analysis and model training.
4. **Handling Outliers:**
   * Identify outliers in numerical features that may adversely affect predictions.
   * Decide on a strategy for handling outliers, such as removing them or transforming the data.
5. **Data Type Conversion:**
   * Ensure that data types are correctly assigned to each feature. For example, dates should be in datetime format, and categorical variables should be encoded appropriately.
6. **Handling Categorical Data:**
   * Encode categorical variables into numerical format using techniques like one-hot encoding or label encoding.
7. **Descriptive Statistics:**
   * Begin with descriptive statistics to summarize and describe the main features of your dataset. Common descriptive statistics include:
     + Measures of central tendency (mean, median, mode)
     + Measures of variability (range, variance, standard deviation)
     + Measures of distribution shape (skewness, kurtosis)
     + Frequency distributions and histograms
8. **Data Visualization:**
   * Create visual representations of the data to aid in understanding. Visualization techniques include histograms, box plots, scatter plots, bar charts, and more.
9. **Inferential Statistics:**
   * Move on to inferential statistics, which involve drawing conclusions from data and making predictions. Key concepts and techniques include:
10. **Statistical Software and Tools:**
    * Utilize statistical software packages (e.g., R, Python with libraries like NumPy, SciPy, and StatsModels) and tools (e.g., Excel, SPSS) to perform analyses.
11. **Interpretation:**
    * Interpret the results of your statistical analysis in the context of your problem or research question. Discuss the practical significance of your findings.
12. **Reporting and Visualization:**
    * Present your results and insights using clear and effective visualizations, tables, and narrative explanations.
    * Use data visualization tools (e.g., Matplotlib, Seaborn, ggplot2) to create informative graphs and charts.

Statistical analysis is a powerful tool for extracting valuable insights from data and making data-driven decisions. It allows you to quantify uncertainty, test hypotheses, and explore relationships within your dataset, ultimately aiding in problem-solving and informed decision-making.

Top of Form

Remember that data cleaning is an iterative process, and it may require revisiting previous steps as you gain more insights into the data or encounter unexpected issues. Regularly checking the quality of your dataset is essential to ensure that your IMDb prediction project is built on a solid foundation.

Top of Form

**Conclusion:**

The IMDb Scores Prediction project seeks to leverage machine learning to enhance our ability to forecast movie ratings accurately. By providing insights into the factors influencing IMDb scores, this project aims to be a valuable resource for the film industry and movie enthusiasts alike.

**Done by,**

Janani A,

Shree Varshenee K