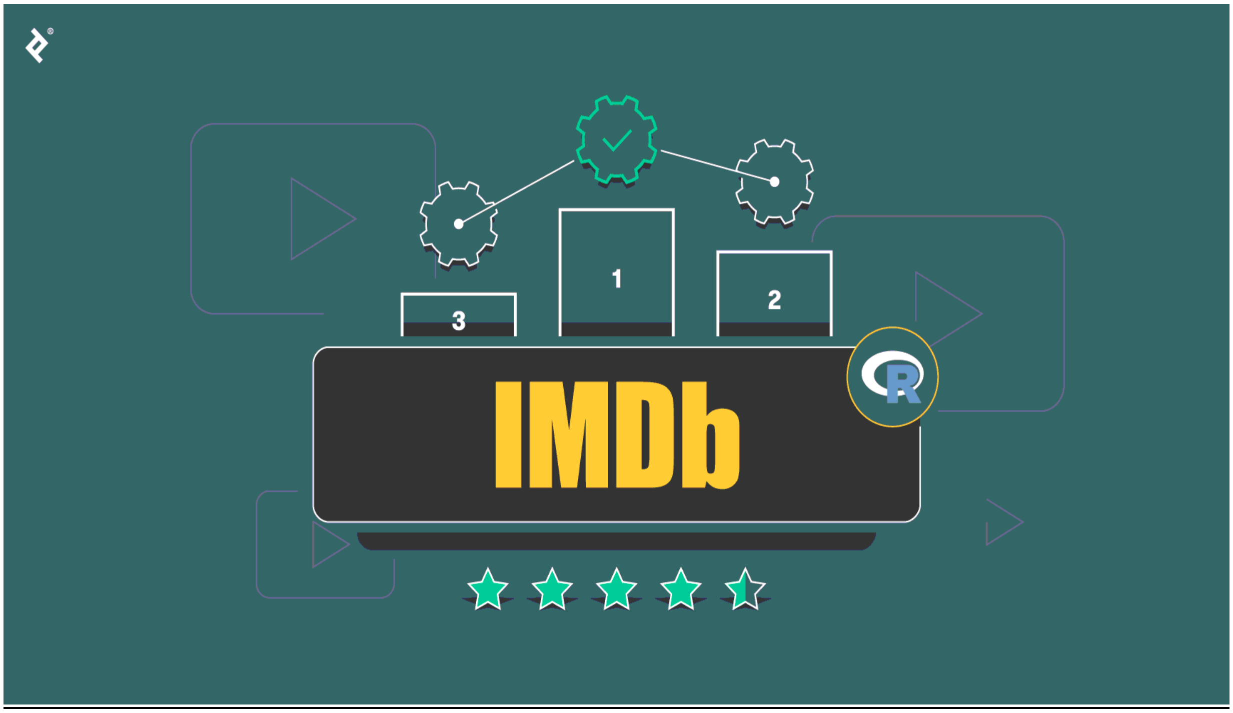
PREDICTING IMDb SCORES

PHASE 2 - INNOVATION



Introduction :

Predicting IMDb scores is a fascinating task that lies at the intersection of data science, machine learning, and the entertainment industry. IMDb, or the Internet Movie Database, is one of the most popular platforms for users to rate and review movies and television shows. Predicting IMDb scores involves developing models that can estimate the average user rating (on a scale of 1 to 10) a movie or TV show might receive based on various features and characteristics. In this introduction, we will provide some background information and a brief data description related to this task.

Background :

IMDb, launched in 1990, is a valuable resource for movie enthusiasts, industry professionals, and critics. It houses a massive database of films and TV shows, along with user-generated ratings and reviews. These ratings provide insights into the popularity and quality of a given production, which can influence the choices of viewers and potential investors.

Predicting IMDb scores has several practical applications. Film studios and producers can use these predictions to gauge the potential success of a movie before its release. Additionally, streaming platforms can use such models to recommend content to users or decide which movies or shows to acquire for their libraries. Researchers and critics can also benefit from these predictions to evaluate the impact and reception of different productions over time.

Data Description :

To predict IMDb scores, you typically work with a dataset that contains a variety of features related to movies or TV shows. These features can include:

1.**Title and Genre :** Information about the title, genre, and subgenre of the production.

2.**Director and Cast :** Details about the director and the cast members.

3.**Production Budget :** The amount of money invested in the production.

4.**Release Date :** The date of release, which can impact a movie's performance.

5.**Runtime :** The duration of the movie or TV show.

6.**Awards and Nominations :** Recognition and accolades received by the production.

7.**User Reviews :** User-generated reviews, comments, and ratings on the IMDb platform.\

8.**Critic Reviews :** Ratings and reviews from professional critics.

9.**Box Office Performance :** Data related to the revenue generated by the production.

10**.Production Company :** Information about the company responsible for making the movie or TV show.

The target variable in this prediction task is the IMDb score, which is typically a numeric value between 1 and 10. Machine learning algorithms are used to build predictive models that can estimate this score based on the provided features. Various regression techniques, such as linear regression, decision trees, or more advanced algorithms like random forests or neural networks, can be employed to create these models.

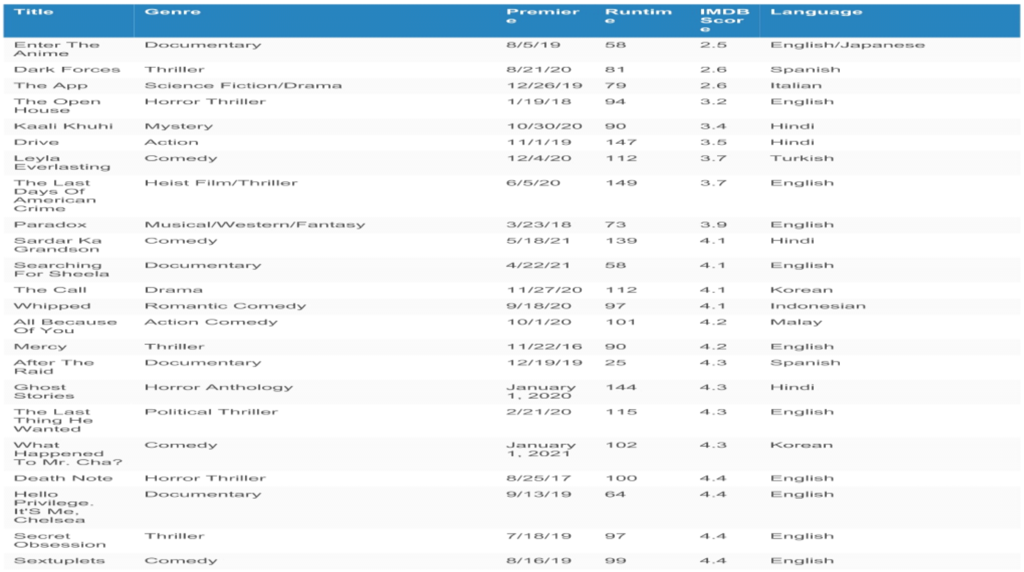
In conclusion, predicting IMDb scores is a valuable application of data analysis and machine learning techniques, offering insights into the film and television industry. The accuracy of these predictions can have a significant impact on investment decisions, content recommendations, and industry trends. It's an exciting field that combines data science and the world of entertainment.

Data Exploration :

Data exploration is a crucial step in any data science or machine learning project, including predicting IMDb scores. It involves understanding and analyzing the dataset to gain insights, identify patterns, and prepare the data for model building.

Data Loading : Start by loading the IMDb dataset into your preferred data analysis environment, such as Python with libraries like Pandas, NumPy, and Matplotlib/Seaborn for visualization.

Dataset Link : (https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores).



Title	Genre	Premiere	Runtime	IMDB Score	Language
Enter The Anime	Documentary	8/5/19	58	2.5	English/Japanese
Dark Forces	Thriller	8/21/20	81	2.6	Spanish
The App	Science Fiction/Drama	12/26/19	79	2.6	Italian
The Open House	Horror Thriller	1/19/18	94	3.2	English
Kaali Khuhi	Mystery	10/30/20	90	3.4	Hindi
Drive	Action	11/1/19	147	3.5	Hindi
Leyla Everlasting	Comedy	12/4/20	112	3.7	Turkish
The Last Days Of American Crime	Heist Film/Thriller	6/5/20	149	3.7	English
Paradox	Musical/Western/Fantasy	3/23/18	73	3.9	English
Sardar Ka Grandson	Comedy	5/18/21	139	4.1	Hindi
Searching For Sheela	Documentary	4/22/21	58	4.1	English
The Call	Drama	11/27/20	112	4.1	Korean
Whipped	Romantic Comedy	9/18/20	97	4.1	Indonesian
All Because Of You	Action Comedy	10/1/20	101	4.2	Malay
Mercy	Thriller	11/22/16	90	4.2	English
After The Raid	Documentary	12/19/19	25	4.3	Spanish
Ghost Stories	Horror Anthology	January 1, 2020	144	4.3	Hindi
The Last Thing He Wanted	Political Thriller	2/21/20	115	4.3	English
What Happened To Mr. Cha?	Comedy	January 1, 2021	102	4.3	Korean
Death Note	Horror Thriller	8/25/17	100	4.4	English
Hello Privilege. It'S Me, Chelsea	Documentary	9/13/19	64	4.4	English
Secret Obsession	Thriller	7/18/19	97	4.4	English
Sextuplets	Comedy	8/16/19	99	4.4	English

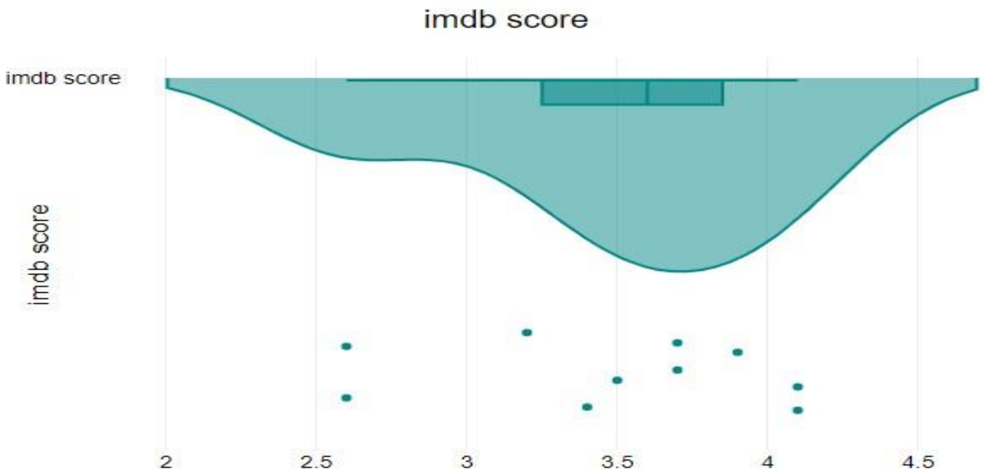
Data cleaning / missing values :

- **Identify Missing Values :** Locate where data is missing in your dataset.
- **Handle Missing Data :** Decide whether to impute (fill in) missing values with means, medians, modes, or other methods, or remove rows/columns with too much missing data.
- **Outliers :** Identify and handle outliers in IMDb scores or other numeric features.
- **Data Type Handling :** Ensure data types are appropriate for analysis.
- **Consistency Checks :** Correct or remove inconsistent or erroneous data.
- **Duplicates :** Check for and remove duplicate entries if present.
- **Scaling/Normalization :** Scale or normalize numeric features if needed for modeling.

Data Visualization :

Data visualization is a crucial step in understanding your data and gaining insights for predicting IMDb scores.

Histogram : Visualize the distribution of IMDb scores to understand whether they are normally distributed or skewed.



Program :

```
# Import necessary libraries
        import pandas as pd
        from sklearn.model_selection import train_test_split
        from sklearn.linear_model import LinearRegression
        from sklearn.metrics import mean_squared_error
         # Load your IMDb dataset (replace 'your_dataset.csv' with the actual file path)
        data = pd.read_csv(' https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores ')
         # Define features (in this example, using 'budget' and 'runtime')
        X = data[['budget', 'runtime']]
        # Define the target variable (IMDb scores)
        y = data['IMDb_Score']
         # Split the data into training and testing sets
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
         # Create a linear regression model
        model = LinearRegression()
        # Train the model on the training data
        model.fit(X_train, y_train)
        # Make predictions on the test data
```

```
y_pred = model.predict(X_test)

# Evaluate the model

mse = mean_squared_error(y_test, y_pred)

print(f'Mean Squared Error: {mse}')

 # You can now use the trained model to predict IMDb scores for new data

# For example, you can predict IMDb scores for a movie with a given budget and runtime

new_data = pd.DataFrame({'budget': [10000000], 'runtime': [120]})

predicted_score = model.predict(new_data)

print(f'Predicted IMDb Score: {predicted_score[0]}')
```

Implement Algorithm :

Supervised Machine Learning Algorithm are used in the Predicting IMDb Scores.

Conclusion :

Predicting IMDb scores involves a multifaceted process that includes data exploration, data cleaning, data visualization, and data preprocessing. Here's a comprehensive conclusion that highlights the key aspects of this predictive task:

Data Exploration:

- Data exploration is the initial step in understanding the IMDb dataset.
- It helps identify patterns, outliers, and relationships in the data.
- Data exploration provides insights into the distribution of IMDb scores, the range of features, and their potential impact on scores.

Data Cleaning:

- Data cleaning is crucial to ensure data quality and reliability for predictions.
- It involves handling missing values, outliers, and inconsistent data.
- Cleaning the dataset helps create a more accurate and reliable basis for modeling.

Data Visualization:

- Data visualization is a powerful tool for understanding the dataset visually.
- Visualizations, such as histograms, scatter plots, and box plots, provide insights into data distributions and correlations.
- Visualizations help identify trends and relationships that may impact IMDb scores.

Data Preprocessing:

- Data preprocessing is essential to prepare the dataset for modeling.
- It includes feature selection, encoding categorical variables, scaling numerical features, and handling target variable transformations.
- Preprocessing ensures that the data is in a suitable format for machine learning algorithms.
- In summary, predicting IMDb scores is a multifaceted task that requires a combination of data exploration, data cleaning, data visualization, and data preprocessing to prepare the dataset and gain insights. These steps set the foundation for building predictive models that can accurately estimate IMDb scores, providing valuable information for various stakeholders in the movie industry and beyond.