In [1]:

```python
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import OneHotEncoder, LabelEncoder
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np
```

In [2]:

```python
#loading dataset of indian movies
dataset_movies ="C:/Users/DELL/Downloads/archive (2)/IMDb Movies India.csv"
movies_df = pd.read_csv(dataset_movies, encoding='latin1')
# Explore the first few rows of the data
print("First few rows of the dataset:")
print(movies_df.head())
```

```
First few rows of the dataset:
                                 Name    Year Duration            Genre  \
0                                         NaN      NaN            Drama
1  #Gadhvi (He thought he was Gandhi) -2019.0  109 min            Drama
2                         #Homecoming -2021.0   90 min   Drama, Musical
3                             #Yaaram -2019.0  110 min  Comedy, Romance
4                   ...And Once Again -2010.0  105 min            Drama

   Rating Votes            Director       Actor 1             Actor 2  \
0     NaN   NaN       J.S. Randhawa      Manmauji              Birbal
1     7.0     8       Gaurav Bakshi  Rasika Dugal      Vivek Ghamande
2     NaN   NaN  Soumyajit Majumdar  Sayani Gupta   Plabita Borthakur
3     4.4    35          Ovais Khan       Prateik          Ishita Raj
4     NaN   NaN        Amol Palekar  Rajat Kapoor  Rituparna Sengupta

           Actor 3
0  Rajendra Bhatia
1    Arvind Jangid
2       Roy Angana
3  Siddhant Kapoor
4      Antara Mali
```

In [3]:

```python
# Explore data types
print("\nData types:")
print(movies_df.dtypes)
```

```
Data types:
Name         object
Year        float64
Duration     object
Genre        object
Rating      float64
Votes        object
Director     object
Actor 1      object
Actor 2      object
Actor 3      object
dtype: object
```

In [4]:

```python
#viewing column names
print(movies_df.columns)
```

```
Index(['Name', 'Year', 'Duration', 'Genre', 'Rating', 'Votes', 'Director',
       'Actor 1', 'Actor 2', 'Actor 3'],
      dtype='object')
```

In [5]:

```python
# Check for missing values
print("\nMissing values:")
print(movies_df.isnull().sum())

# Handle missing values (e.g., fill with mean, median, or mode)
movies_df.fillna(0, inplace=True)
```

```
Missing values:
Name           0
Year         528
Duration    8269
Genre       1877
Rating      7590
Votes       7589
Director     525
Actor 1     1617
Actor 2     2384
Actor 3     3144
dtype: int64
```

In [6]:

```python
# Check for missing values
print("\nMissing values:")
print(movies_df.isnull().sum())
```

```
Missing values:
Name        0
Year        0
Duration    0
Genre       0
Rating      0
Votes       0
Director    0
Actor 1     0
Actor 2     0
Actor 3     0
dtype: int64
```

In [7]:

```python
# Data Cleaning
# Convert 'Year' and 'Duration' columns to appropriate data types
movies_df['Year'] = movies_df['Year'].astype(int)
movies_df['Duration'] = movies_df['Duration'].str.extract('(\d+)').astype(float)
movies_df.fillna(0, inplace=True)
movies_df['Duration'] = movies_df['Duration'].astype(int)
print(movies_df.head())
```

```
                                 Name  Year  Duration            Genre  \
0                                         0         0            Drama
1  #Gadhvi (He thought he was Gandhi) -2019       109            Drama
2                         #Homecoming -2021        90   Drama, Musical
3                             #Yaaram -2019       110  Comedy, Romance
4                   ...And Once Again -2010       105            Drama

   Rating Votes            Director       Actor 1             Actor 2  \
0     0.0     0       J.S. Randhawa      Manmauji              Birbal
1     7.0     8       Gaurav Bakshi  Rasika Dugal      Vivek Ghamande
2     0.0     0  Soumyajit Majumdar  Sayani Gupta   Plabita Borthakur
3     4.4    35          Ovais Khan       Prateik          Ishita Raj
4     0.0     0        Amol Palekar  Rajat Kapoor  Rituparna Sengupta

           Actor 3
0  Rajendra Bhatia
1    Arvind Jangid
2       Roy Angana
3  Siddhant Kapoor
4      Antara Mali
```

In [8]:

```python
# Explore data types
print("\nData types:")
print(movies_df.dtypes)
```

```
Data types:
Name         object
Year          int32
Duration      int32
Genre        object
Rating      float64
Votes        object
Director     object
Actor 1      object
Actor 2      object
Actor 3      object
dtype: object
```

In [9]:

```python
# Convert the 'Votes' column to string type
movies_df['Votes'] = movies_df['Votes'].astype(str)
# Remove commas and replace 'nan' with '0' in the 'Votes' column
movies_df['Votes'] = movies_df['Votes'].str.replace(',', '').str.replace('nan', '0')
movies_df['Votes'] = movies_df['Votes'].str.extract('(\d+)').astype(float)

# Convert the 'Votes' column to float first to handle values like '8.0'
movies_df['Votes'] = movies_df['Votes'].astype(float)

#Convert the 'Votes' column to integer
movies_df['Votes'] = movies_df['Votes'].astype(int)

# Check the data type of 'Votes' column
print("Data type of 'Votes' column:", movies_df['Votes'].dtype)
```

```
Data type of 'Votes' column: int32
```

In [10]:

```python
print(movies_df.head())
```

```
                                 Name  Year  Duration            Genre  \
0                                         0         0            Drama
1  #Gadhvi (He thought he was Gandhi) -2019       109            Drama
2                         #Homecoming -2021        90   Drama, Musical
3                             #Yaaram -2019       110  Comedy, Romance
4                   ...And Once Again -2010       105            Drama

   Rating  Votes            Director       Actor 1             Actor 2  \
0     0.0      0       J.S. Randhawa      Manmauji              Birbal
1     7.0      8       Gaurav Bakshi  Rasika Dugal      Vivek Ghamande
2     0.0      0  Soumyajit Majumdar  Sayani Gupta   Plabita Borthakur
3     4.4     35          Ovais Khan       Prateik          Ishita Raj
4     0.0      0        Amol Palekar  Rajat Kapoor  Rituparna Sengupta

           Actor 3
0  Rajendra Bhatia
1    Arvind Jangid
2       Roy Angana
3  Siddhant Kapoor
4      Antara Mali
```

In [11]:

```python
print(movies_df['Genre'].value_counts())
```

```
Drama                          2780
0                              1877
Action                         1289
Thriller                        779
Romance                         708
                               ...
Action, Musical, War              1
Horror, Crime, Thriller           1
Animation, Comedy                 1
Romance, Action, Crime            1
Adventure, Fantasy, Sci-Fi        1
Name: Genre, Length: 486, dtype: int64
```

In [ ]:

In [ ]:

In [12]:

```python
top_n = 5

# Define columns with high cardinality to be encoded
high_cardinality_columns = ['Actor 1', 'Actor 2', 'Actor 3', 'Genre']

# Apply one-hot encoding for each high-cardinality column
for column in high_cardinality_columns:
    # Calculate the frequency of each category
    category_counts = movies_df[column].value_counts()

    # Identify the top N most frequent categories
    top_categories = category_counts.nlargest(top_n).index

    # Create a new column in the DataFrame with one-hot encoding for each top category
    for category in top_categories:
        movies_df[f"{column}_{category}"] = (movies_df[column] == category).astype(int)

    # Group all other categories as 'Other'
    movies_df[f"{column}_Other"] = ~movies_df[column].isin(top_categories).astype(int)

    # Drop the original column if needed
    # df.drop(column, axis=1, inplace=True)

# Display the updated DataFrame
print(movies_df)
```

```
                                     Name  Year  Duration            Genre  \
0                                             0         0            Drama
1      #Gadhvi (He thought he was Gandhi) -2019       109            Drama
2                             #Homecoming -2021        90   Drama, Musical
3                                 #Yaaram -2019       110  Comedy, Romance
4                       ...And Once Again -2010       105            Drama
...                                   ...   ...       ...              ...
15504                 Zulm Ko Jala Doonga -1988         0           Action
15505                               Zulmi -1999       129    Action, Drama
15506                           Zulmi Raj -2005         0           Action
15507                       Zulmi Shikari -1988         0           Action
15508                        Zulm-O-Sitam -1998       130    Action, Drama

       Rating  Votes            Director           Actor 1  \
0         0.0      0       J.S. Randhawa          Manmauji
1         7.0      8       Gaurav Bakshi      Rasika Dugal
2         0.0      0  Soumyajit Majumdar      Sayani Gupta
3         4.4     35          Ovais Khan           Prateik
4         0.0      0        Amol Palekar      Rajat Kapoor
...       ...    ...                 ...               ...
15504     4.6     11       Mahendra Shah  Naseeruddin Shah
15505     4.5    655          Kuku Kohli      Akshay Kumar
15506     0.0      0          Kiran Thej   Sangeeta Tiwari
15507     0.0      0                   0                 0
15508     6.2     20        K.C. Bokadia        Dharmendra

                  Actor 2          Actor 3  ...  Actor 3_Shakti Kapoor  \
0                  Birbal  Rajendra Bhatia  ...                      0
1          Vivek Ghamande    Arvind Jangid  ...                      0
2       Plabita Borthakur       Roy Angana  ...                      0
3              Ishita Raj  Siddhant Kapoor  ...                      0
4      Rituparna Sengupta      Antara Mali  ...                      0
...                   ...              ...  ...                    ...
15504       Sumeet Saigal    Suparna Anand  ...                      0
15505      Twinkle Khanna      Aruna Irani  ...                      0
15506                   0                0  ...                      0
15507                   0                0  ...                      0
15508          Jaya Prada      Arjun Sarja  ...                      0

       Actor 3_Anupam Kher  Actor 3_Jeevan  Actor 3_Other  Genre_Drama  \
0                        0               0             -1            1
1                        0               0             -1            1
2                        0               0             -1            0
3                        0               0             -1            0
4                        0               0             -1            1
...                    ...             ...            ...          ...
15504                    0               0             -1            0
15505                    0               0             -1            0
15506                    0               0             -2            0
15507                    0               0             -2            0
15508                    0               0             -1            0

       Genre_0  Genre_Action  Genre_Thriller  Genre_Romance  Genre_Other
0            0             0               0              0           -2
1            0             0               0              0           -2
2            0             0               0              0           -1
3            0             0               0              0           -1
4            0             0               0              0           -2
...        ...           ...             ...            ...          ...
15504        0             1               0              0           -2
15505        0             0               0              0           -1
15506        0             1               0              0           -2
15507        0             1               0              0           -2
15508        0             0               0              0           -1

[15509 rows x 34 columns]
```

In [13]:

```python
df = pd.DataFrame(movies_df)


# List of columns to remove
columns_to_remove = ['Name', 'Genre','Director','Actor 1','Actor 2','Actor 3']

# Remove columns from the DataFrame
moviedf=df.drop(columns=columns_to_remove)
moviedf=pd.DataFrame(moviedf)

# Display the DataFrame after removing columns
print(moviedf)
```

```
       Year  Duration  Rating  Votes  Actor 1_0  Actor 1_Ashok Kumar  \
0         0         0     0.0      0          0                    0
1     -2019       109     7.0      8          0                    0
2     -2021        90     0.0      0          0                    0
3     -2019       110     4.4     35          0                    0
4     -2010       105     0.0      0          0                    0
...     ...       ...     ...    ...        ...                  ...
15504 -1988         0     4.6     11          0                    0
15505 -1999       129     4.5    655          0                    0
15506 -2005         0     0.0      0          0                    0
15507 -1988         0     0.0      0          1                    0
15508 -1998       130     6.2     20          0                    0

       Actor 1_Jeetendra  Actor 1_Dharmendra  Actor 1_Mithun Chakraborty  \
0                      0                   0                           0
1                      0                   0                           0
2                      0                   0                           0
3                      0                   0                           0
4                      0                   0                           0
...                  ...                 ...                         ...
15504                  0                   0                           0
15505                  0                   0                           0
15506                  0                   0                           0
15507                  0                   0                           0
15508                  0                   1                           0

       Actor 1_Other  ...  Actor 3_Shakti Kapoor  Actor 3_Anupam Kher  \
0                 -1  ...                      0                    0
1                 -1  ...                      0                    0
2                 -1  ...                      0                    0
3                 -1  ...                      0                    0
4                 -1  ...                      0                    0
...              ...  ...                    ...                  ...
15504             -1  ...                      0                    0
15505             -1  ...                      0                    0
15506             -1  ...                      0                    0
15507             -2  ...                      0                    0
15508             -2  ...                      0                    0

       Actor 3_Jeevan  Actor 3_Other  Genre_Drama  Genre_0  Genre_Action  \
0                   0             -1            1        0             0
1                   0             -1            1        0             0
2                   0             -1            0        0             0
3                   0             -1            0        0             0
4                   0             -1            1        0             0
...               ...            ...          ...      ...           ...
15504               0             -1            0        0             1
15505               0             -1            0        0             0
15506               0             -2            0        0             1
15507               0             -2            0        0             1
15508               0             -1            0        0             0

       Genre_Thriller  Genre_Romance  Genre_Other
0                   0              0           -2
1                   0              0           -2
2                   0              0           -1
3                   0              0           -1
4                   0              0           -2
...               ...            ...          ...
15504               0              0           -2
15505               0              0           -1
15506               0              0           -2
15507               0              0           -2
15508               0              0           -1

[15509 rows x 28 columns]
```

In [14]:

```python
X = moviedf.drop('Rating', axis=1)
y = moviedf['Rating']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Choose a model and train it
model = RandomForestRegressor()
model.fit(X_train, y_train)
```

Out[14]:

```
▾ RandomForestRegressor
RandomForestRegressor()
```

In [15]:

```python
y_pred = model.predict(X_test)

# Calculate RMSE and R-squared
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

print(f'RMSE: {rmse}')
print(f'R-squared: {r2}')
```

```
RMSE: 0.846337140566508
R-squared: 0.9234600155275023
```

In [16]:

```
Best parameters: {'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 150}
RandomForestRegressor(max_depth=10, min_samples_split=10, n_estimators=150)
```

In [ ]: