CODE

```
# Importing Libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error
import matplotlib.pyplot as plt
import seaborn as sns

# Step 1: Load Dataset
# Assuming the dataset is saved as 'movies.csv'
df = pd.read_csv('imbd movies.csv', encoding='ISO-8859-1')

print(df)

# Step 2: Exploratory Data Analysis (EDA)
print(df.head())
print(df.info())
print(df.describe())

# Visualizing the target variable
sns.histplot(df['Rating'], kde=True)
plt.title('Distribution of Movie Ratings')
plt.show()
```

```
# Step 3: Data Preprocessing
# Handling missing values
df.fillna({'Genre': 'Unknown', 'Director': 'Unknown', 'Actors': 'Unknown'}, inplace=True)
df['Budget'] = 0  # Assigning a default value of 0 to all rows
print(df.fillna)
print(df['Budget'])

# Splitting data into features and target
print(df.columns)
columns_to_select = ['Budget', 'Year', 'Genre', 'Director', 'Actor']
existing_columns = [col for col in columns_to_select if col in df.columns]
X = df[existing_columns]

columns = ['Budget', 'Year', 'Genre', 'Director', 'Actor']
X = df[[col for col in columns if col in df.columns]]
print(X)

from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder

# Define columns for numerical and categorical preprocessing
numerical_features = ['Budget', 'Year']
categorical_features = ['Genre', 'Director']

# Define preprocessing steps
numerical_transformer = StandardScaler()
categorical_transformer = OneHotEncoder(handle_unknown='ignore')
```

```
# Combine preprocessing steps
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, numerical_features),
        ('cat', categorical_transformer, categorical_features)
    ]
)
X_preprocessed = preprocessor.fit_transform(X)
print(X_preprocessed)

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression

# Example dataset (replace with your actual data loading)
df = pd.DataFrame({
    'feature1': [1, 2, 3, 4, 5],
    'feature2': [5, 4, 3, 2, 1],
    'target': [0, 1, 0, 1, 0]
})

# Split features and target
X = df[['feature1', 'feature2']]
y = df['target']

# Split 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)

# Define and train the model
```

```
model = LogisticRegression()
model.fit(X_train, y_train)

# Step 6: Feature Importance

import pandas as pd
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.ensemble import RandomForestClassifier
import seaborn as sns
import matplotlib.pyplot as plt

# Example features
categorical_features = ['cat_feature1', 'cat_feature2']
numerical_features = ['num_feature1', 'num_feature2']

# Preprocessor
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numerical_features),
        ('cat', OneHotEncoder(), categorical_features)
    ]
)

# Define pipeline
model = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('regressor', RandomForestClassifier())
```

```
])

# Example data
X_train = pd.DataFrame({
    'cat_feature1': ['A', 'B', 'A', 'B'],
    'cat_feature2': ['X', 'Y', 'X', 'Y'],
    'num_feature1': [1.0, 2.0, 3.0, 4.0],
    'num_feature2': [5.0, 6.0, 7.0, 8.0]
})
y_train = [0, 1, 0, 1]

# Train the pipeline
model.fit(X_train, y_train)

# Access the regressor and feature importance
regressor = model.named_steps['regressor']
if hasattr(regressor, 'feature_importances_'):
    # Get feature names
    feature_names = numerical_features +
list(preprocessor.named_transformers_['cat'].get_feature_names_out(categorical_features))

    # Get feature importances
    feature_importances = regressor.feature_importances_

    # Plot feature importance
    importance_df = pd.DataFrame({
        'Feature': feature_names,
        'Importance': feature_importances
    }).sort_values(by='Importance', ascending=False)
```

```
    sns.barplot(x='Importance', y='Feature', data=importance_df.head(10))
    plt.title('Top 10 Feature Importances')
    plt.show()
else:
    print("The regressor does not have feature_importances_ attribute.")


# Step 7: Predict for new data
import pandas as pd
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.ensemble import RandomForestClassifier

# Define features
categorical_features = ['cat_feature1', 'cat_feature2']
numerical_features = ['num_feature1', 'num_feature2']

# Preprocessor
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numerical_features),
        ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features)
    ]
)

# Define pipeline
model = Pipeline(steps=[
```

```
    ('preprocessor', preprocessor),
    ('regressor', RandomForestClassifier())
])

# Example training data
X_train = pd.DataFrame({
    'cat_feature1': ['A', 'B', 'A', 'B'],
    'cat_feature2': ['X', 'Y', 'X', 'Y'],
    'num_feature1': [1.0, 2.0, 3.0, 4.0],
    'num_feature2': [5.0, 6.0, 7.0, 8.0]
})
y_train = [0, 1, 0, 1]

# Train the pipeline
model.fit(X_train, y_train)

# Example new data for prediction
new_data = pd.DataFrame({
    'cat_feature1': ['A'],
    'cat_feature2': ['X'],
    'num_feature1': [1.5],
    'num_feature2': [6.5]
})

# Ensure all columns are present
for col in ['cat_feature1', 'cat_feature2', 'num_feature1', 'num_feature2']:
    if col not in new_data:
        new_data[col] = None
# Predict
```

```
predicted_rating = model.predict(new_data)

print(predicted_rating)
```

OUTPUT

```
                Name    Year Duration          Genre  ...          Director         Actor 1          Actor 2          Actor
3

0                              NaN      NaN          Drama  ...      J.S. Randhawa         Manmauji            Birbal
Rajendra Bhatia

1      #Gadhvi (He thought he was Gandhi) -2019.0  109 min          Drama  ...      Gaurav Bakshi      Rasika
Dugal      Vivek Ghamande    Arvind Jangid

2                    #Homecoming -2021.0   90 min   Drama, Musical  ...  Soumyajit Majumdar      Sayani
Gupta   Plabita Borthakur       Roy Angana

3                     #Yaaram -2019.0  110 min  Comedy, Romance  ...         Ovais Khan          Prateik
Ishita Raj  Siddhant Kapoor

4                  ...And Once Again -2010.0  105 min          Drama  ...      Amol Palekar      Rajat Kapoor
Rituparna Sengupta      Antara Mali

...                          ...    ...      ...           ...  ...              ...              ...              ...
...

15504            Zulm Ko Jala Doonga -1988.0      NaN         Action  ...      Mahendra Shah  Naseeruddin
Shah      Sumeet Saigal    Suparna Anand

15505                  Zulmi -1999.0  129 min    Action, Drama  ...         Kuku Kohli      Akshay Kumar
Twinkle Khanna      Aruna Irani

15506               Zulmi Raj -2005.0      NaN         Action  ...         Kiran Thej   Sangeeta Tiwari
NaN          NaN

15507             Zulmi Shikari -1988.0      NaN         Action  ...              NaN              NaN              NaN
NaN

15508            Zulm-O-Sitam -1998.0  130 min    Action, Drama  ...       K.C. Bokadia        Dharmendra
Jaya Prada      Arjun Sarja


[15509 rows x 10 columns]

                    Name    Year Duration          Genre  Rating Votes          Director       Actor 1
Actor 2         Actor 3
```

0 NaN NaN Drama NaN NaN J.S. Randhawa Manmauji Birbal Rajendra Bhatia

1 #Gadhvi (He thought he was Gandhi) -2019.0 109 min Drama 7.0 8 Gaurav Bakshi Rasika Dugal Vivek Ghamande Arvind Jangid

2 #Homecoming -2021.0 90 min Drama, Musical NaN NaN Soumyajit Majumdar Sayani Gupta Plabita Borthakur Roy Angana

3 #Yaaram -2019.0 110 min Comedy, Romance 4.4 35 Ovais Khan Prateik Ishita Raj Siddhant Kapoor

4 ...And Once Again -2010.0 105 min Drama NaN NaN Amol Palekar Rajat Kapoor Rituparna Sengupta Antara Mali

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 15509 entries, 0 to 15508

Data columns (total 10 columns):

#	Column	Non-Null Count	Dtype
0	Name	15509 non-null	object
1	Year	14981 non-null	float64
2	Duration	7240 non-null	object
3	Genre	13632 non-null	object
4	Rating	7919 non-null	float64
5	Votes	7920 non-null	object
6	Director	14984 non-null	object
7	Actor 1	13892 non-null	object
8	Actor 2	13125 non-null	object
9	Actor 3	12365 non-null	object

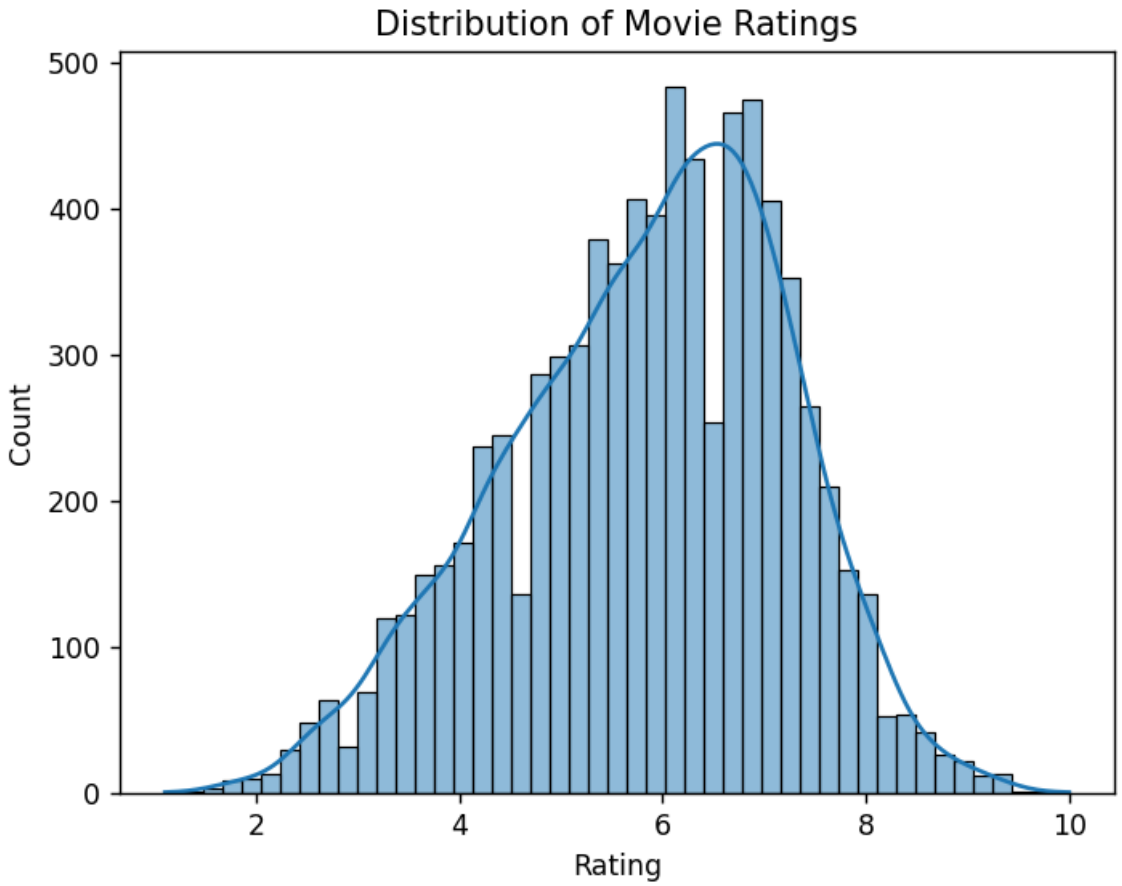
dtypes: float64(2), object(8)

memory usage: 1.2+ MB

None

	Year	Rating
count	14981.000000	7919.000000

mean	-1987.012215	5.841621
std	25.416689	1.381777
min	-2022.000000	1.100000
25%	-2009.000000	4.900000
50%	-1991.000000	6.000000
75%	-1968.000000	6.800000
max	-1913.000000	10.000000



<bound method NDFrame.fillna of Name Year Duration Genre ... Actor 1 Actor 2 Actor 3 Budget

0 NaN NaN Drama ... Manmauji Birbal Rajendra Bhatia 0

1 #Gadhvi (He thought he was Gandhi) -2019.0 109 min Drama ... Rasika Dugal Vivek Ghamande Arvind Jangid 0

2 #Homecoming -2021.0 90 min Drama, Musical ... Sayani Gupta Plabita Borthakur Roy Angana 0

3 #Yaaram -2019.0 110 min Comedy, Romance ... Prateik Ishita Raj Siddhant Kapoor 0

4 ...And Once Again -2010.0 105 min Drama ... Rajat Kapoor Rituparna Sengupta Antara Mali 0

... ...

15504 Zulm Ko Jala Doonga -1988.0 NaN Action ... Naseeruddin Shah Sumeet Saigal Suparna Anand 0

15505 Zulmi -1999.0 129 min Action, Drama ... Akshay Kumar Twinkle Khanna Aruna Irani 0

15506 Zulmi Raj -2005.0 NaN Action ... Sangeeta Tiwari NaN NaN 0

15507 Zulmi Shikari -1988.0 NaN Action ... NaN NaN NaN 0

15508 Zulm-O-Sitam -1998.0 130 min Action, Drama ... Dharmendra Jaya Prada Arjun Sarja 0

[15509 rows x 11 columns]>

0 0

1 0

2 0

3 0

4 0

..

15504 0

15505 0

15506 0

15507 0

15508 0

Name: Budget, Length: 15509, dtype: int64

Index(['Name', 'Year', 'Duration', 'Genre', 'Rating', 'Votes', 'Director',

'Actor 1', 'Actor 2', 'Actor 3', 'Budget'],

dtype='object')

	Budget	Year	Genre	Director
0	0	NaN	Drama	J.S. Randhawa
1	0	-2019.0	Drama	Gaurav Bakshi
2	0	-2021.0	Drama, Musical	Soumyajit Majumdar
3	0	-2019.0	Comedy, Romance	Ovais Khan
4	0	-2010.0	Drama	Amol Palekar
...	...	...	...	...
15504	0	-1988.0	Action	Mahendra Shah
15505	0	-1999.0	Action, Drama	Kuku Kohli
15506	0	-2005.0	Action	Kiran Thej
15507	0	-1988.0	Action	Unknown
15508	0	-1998.0	Action, Drama	K.C. Bokadia

[15509 rows x 4 columns]

<Compressed Sparse Row sparse matrix of dtype 'float64'

with 46527 stored elements and shape (15509, 6427)>

Coords	Values
(0, 1)	nan
(0, 301)	1.0
(0, 2414)	1.0
(1, 1)	-1.2585767028126105
(1, 301)	1.0
(1, 2036)	1.0
(2, 1)	-1.337267785763205
(2, 353)	1.0

(2, 5611) 1.0

(3, 1) -1.2585767028126105

(3, 230) 1.0

(3, 3807) 1.0

(4, 1) -0.9044668295349345

(4, 301) 1.0

(4, 873) 1.0

(5, 1) -0.39297479035606947

(5, 199) 1.0

(5, 4288) 1.0

(6, 1) -0.707739122158448

(6, 368) 1.0

(6, 5481) 1.0

(7, 1) -0.8257757465843399

(7, 264) 1.0

(7, 963) 1.0

(8, 1) -0.9831579124855292

: :

(15500, 1337) 1.0

(15501, 1) -0.1962470829795829

(15501, 31) 1.0

(15501, 1433) 1.0

(15502, 1) 0.31524495619928217

(15502, 2) 1.0

(15502, 1886) 1.0

(15503, 1) -0.07821045855369096

(15503, 31) 1.0

(15503, 4904) 1.0

(15504, 1) -0.03886491707839365

(15504, 2) 1.0

(15504, 3178) 1.0

(15505, 1) -0.4716658733066641

(15505, 42) 1.0

(15505, 2987) 1.0

(15506, 1) -0.707739122158448

(15506, 2) 1.0

(15506, 2912) 1.0

(15507, 1) -0.03886491707839365

(15507, 2) 1.0

(15507, 6037) 1.0

(15508, 1) -0.4323203318313668

(15508, 42) 1.0

(15508, 2683) 1.0