IMDb Score Prediction using Data Science…

TEAM MEMBER : KARUNYA L S
NAAN MUTHALVAN ID : au723721243026

Introduction:

Problem Definition:IMDb scores are determined by user ratings and can change over time as more users rate the movie or show.

The problem is to develop a machine learning model to predict the IMDb scores of movies available on Films based on their genre, premiere date, runtime, and language.

This project involves data collection, data prepossessing, feature engineering, clustering algorithms, visualization, and interpretation of results.

The model aims to accurately estimate the popularity of movies to assist users in discovering highly rated films that align with their preferences.

Project Phase 2: Develop a machine learning model to predict the IMDb scores of movies available on Films based on their genre, premiere date, runtime, and language.

Dataset:

Datasetlink :https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores/

Title	Genre	Premiere	Runtime	IMDB Score	Language
Enter the A	Documenta	5-Aug-19	58	2.5	English/Japanese
Dark Forces	Thriller	21-Aug-20	81	2.6	Spanish
The App	Science fict	26-Dec-19	79	2.6	Italian
The Open H	Horror thril	19-Jan-18	94	3.2	English
Kaali Khuhi	Mystery	30-Oct-20	90	3.4	Hindi
Drive	Action	1-Nov-19	147	3.5	Hindi
Leyla Everla	Comedy	4-Dec-20	112	3.7	Turkish
The Last Da	Heist film/	5-Jun-20	149	3.7	English
Paradox	Musical/We	23-Mar-18	73	3.9	English
Sardar Ka G	Comedy	18-May-21	139	4.1	Hindi
Searching f	Documenta	22-Apr-21	58	4.1	English
The Call	Drama	27-Nov-20	112	4.1	Korean
Whipped	Romantic c	18-Sep-20	97	4.1	Indonesian
All Because	Action com	1-Oct-20	101	4.2	Malay
Mercy	Thriller	22-Nov-16	90	4.2	English
After the R	Documenta	19-Dec-19	25	4.3	Spanish
Ghost Stori	Horror anth	1-Jan-20	144	4.3	Hindi
The Last Th	Political thr	21-Feb-20	115	4.3	English
What Happ	Comedy	1-Jan-21	102	4.3	Korean
Death Note	Horror thril	25-Aug-17	100	4.4	English
Hello Privil	Documenta	13-Sep-19	64	4.4	English

The Girl on	Thriller	26-Feb-21	120	4.4	Hindi
Thunder Fo	Superhero-	9-Apr-21	105	4.4	English
Fatal Affair	Thriller	16-Jul-20	89	4.5	English
Just Say Ye	Romantic c	2-Apr-21	97	4.5	Dutch
Seriously Si	Comedy	31-Jul-20	107	4.5	English
The Misadv	Comedy	10-Feb-21	99	4.5	French
5 Star Chris	Comedy	7-Dec-18	95	4.6	Italian
After Maria	Documenta	24-May-19	37	4.6	English/Spanish
I Am the Pr	Horror	28-Oct-16	89	4.6	English
Paris Is Us	Romance d	22-Feb-19	83	4.6	French
Porta dos F	Comedy	3-Dec-19	46	4.6	Portuguese
Rattlesnake	Horror	25-Oct-19	85	4.6	English
The Players	Comedy	15-Jul-20	88	4.6	Italian
We Are On	Documenta	14-Jul-20	86	4.6	French
Finding Agr	Drama	30-Nov-20	105	4.7	Filipino
IO	Science fict	18-Jan-19	95	4.7	English
Sentinelle	Action	5-Mar-21	80	4.7	French
Sol Levante	Anime / Sh	2-Apr-20	4	4.7	English
The Bindin	Drama	2-Oct-20	93	4.7	Italian
We Can Be	Superhero	25-Dec-20	100	4.7	English
Christmas (	Thriller	4-Dec-20	106	4.8	German
Coin Heist	Heist	6-Jan-17	97	4.8	English

Program :

```
import numpy as np
import pandas as pd
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from datetime import datetime, timedelta
ds = pd.read_csv("/kaggle/input/netflix-original
films-imdb-scores/NetflixOriginals.csv",encoding
= "ISO-8859-1")
ds_date = ds.copy()
ds.head(5)
```

	Title	Genre	Premiere	Runtime	IMDB Score	Language
0	Enter the Anime	Documentary	August 5, 2019	58	2.5	English/Japanese
1	Dark Forces	Thriller	August 21, 2020	81	2.6	Spanish
2	The App	Science fiction/Drama	December 26, 2019	79	2.6	Italian
3	The Open House	Horror thriller	January 19, 2018	94	3.2	English
4	Kaali Khuhi	Mystery	October 30, 2020	90	3.4	Hindi

```
ds.describe().T
ds.info(verbose=True,show_counts=True)
ds.isna().sum()
ds['Title'].value_counts()
ds['Genre'].value_counts()
ds['Premiere'].value_counts()
```

```
ds_date["Premiere"]=ds_date["Premiere"].apply(la
mbda x: "".join(x for x in x.replace(".",",")))
ds_date["PremiereDate"]=ds_date["Premiere"].appl
y(lambda x: datetime.strptime(x,
"%B %d, %Y").date())
ds_date["Year"] =
ds_date["Premiere"].apply(lambda x: "".join(x for
x in x.replace(",","").split()[-1]))
ds_date["PremiereDate"] =
pd.to_datetime(ds_date["PremiereDate"])
ds_date
ds_date.info()
ds['Language'].value_counts()
ds['Genre'].value_counts()
genre = ds['Genre'].value_counts()
genre.head()
plt.figure(figsize=(16, 5))
ds['Genre'].value_counts().head(10).plot(kind='bar',
color='red')
plt.xlabel('Genre')
plt.ylabel('Number of Genre')
plt.xticks(rotation=90)
plt.show(block=True)
ds['Language'].value_counts()
ds_lang = ds['Language'].value_counts()
ds_lang.head(5).plot(kind='bar', color='red')
plt.xlabel('Languge')
plt.ylabel('Number of Language')
plt.show(block=True)
```

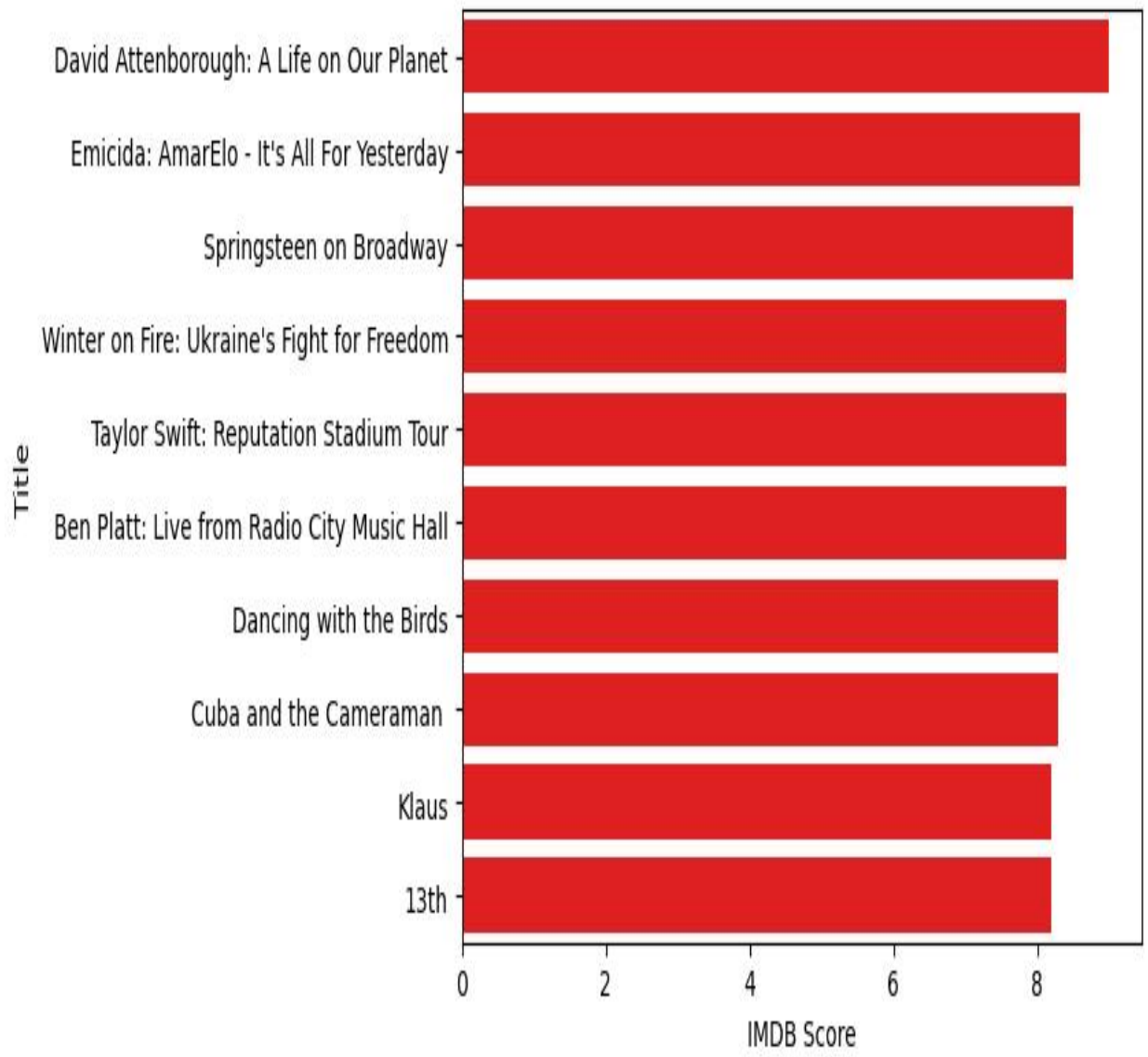
```
ds.groupby('Language').agg({'Runtime':
'sum'}).sort_values('Runtime',
ascending=False).head(5).plot(kind='bar',color='red
')
plt.xlabel('Language')
plt.ylabel('Runtime')
plt.show(block=True)
ds_english =
ds[ds['Language']=='English'].sort_values('IMDB
Score', ascending=False)
ds_english.head()
ds_date.groupby('Year').agg({'Runtime':
'sum'}).sort_values('Runtime',
ascending=False).plot(kind='bar', color='red')
plt.xlabel('Year')
plt.ylabel('Sum of Runtime')
plt.show(block=True)
ds_date.groupby('Year').agg({'Title':
'count'}).sort_values('Title',
ascending=False).plot(kind='bar', color='red')
plt.xlabel('Year')
plt.ylabel('Number of Film')
plt.show(block=True)
```

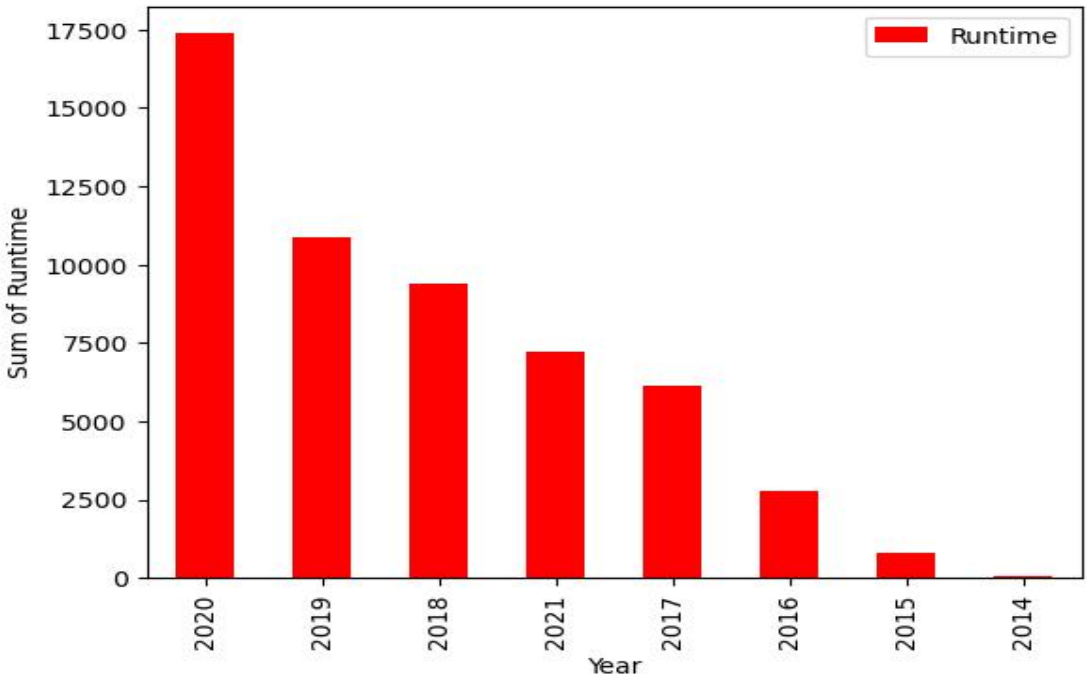
Output:

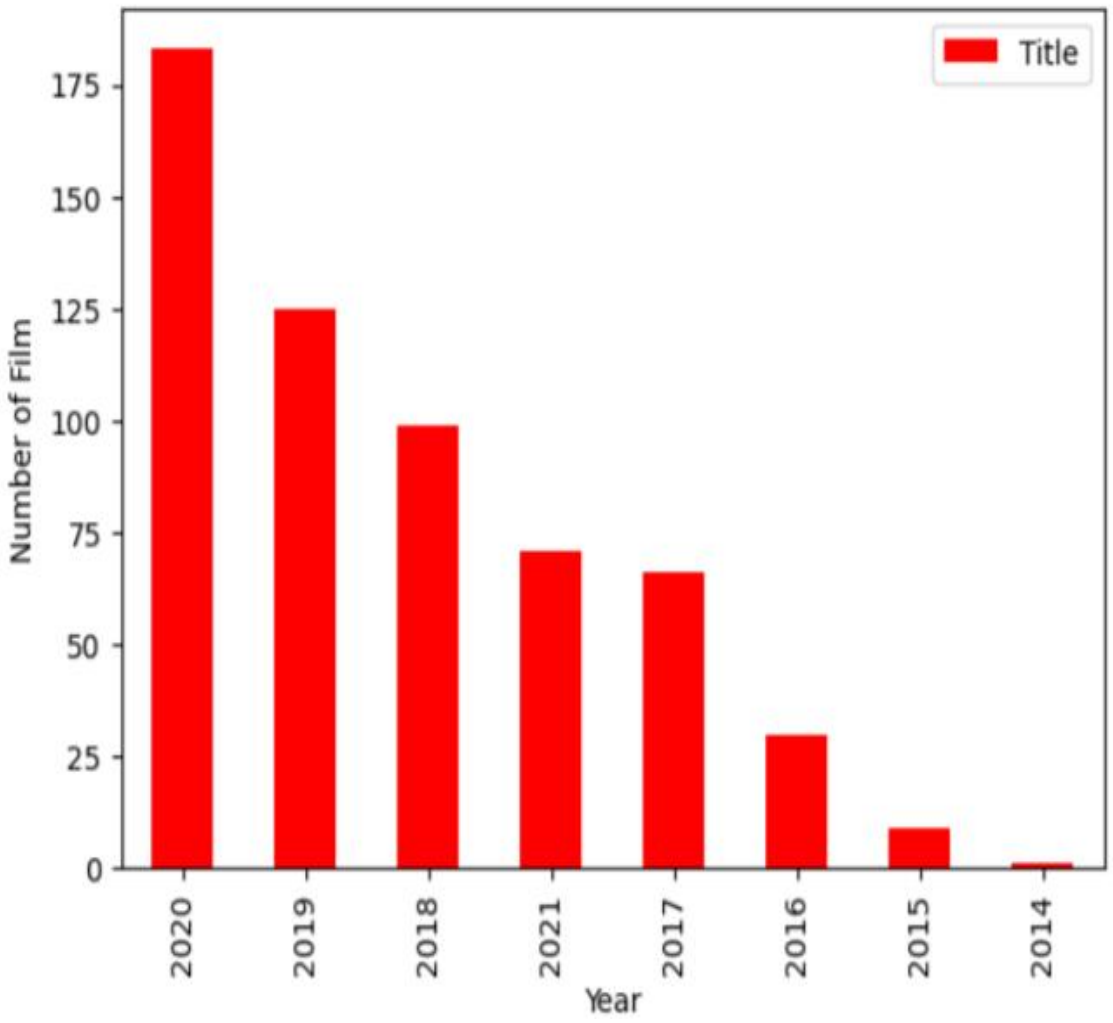
```
English          401
Hindi             33
Spanish            31
```

French	20
Italian	14
Portuguese	12
Indonesian	9
Japanese	6
Korean	6
German	5
Turkish	5
English/Spanish	5
Polish	3
Dutch	3
Marathi	3
English/Hindi	2
Thai	2
English/Mandarin	2
English/Japanese	2
Filipino	2
English/Russian	1
Bengali	1
English/Arabic	1
English/Korean	1
Spanish/English	1
Tamil	1
English/Akan	1
Khmer/English/French	1
Swedish	1
Georgian	1
Thia/English	1
English/Taiwanese/Mandarin	1
English/Swedish	1

Spanish/Catalan 1
Spanish/Basque 1
Norwegian 1
Malay 1
English/Ukranian/Russian 1
Name: Language, dtype: int64




Runtime
17500
15000
12500
10000
7500
5000
2500
0
Sum of Runtime
2020
2019
2018
2021
2017
2016
2015
2014
Year


Title
175
150
125
100
75
50
25
0
Number of Film
2020
2019
2018
2021
2017
2016
2015
2014
Year

Conclusion :

In conclusion, predicting IMDb scores is a complex task that involves various factors and challenges.IMDb scores are influenced by a multitude of subjective and contextual factors, and no model can perfectly capture all of these nuances.

To improve IMDb score predictions, it's crucial to consider factors such as user reviews, genre, director, actors, and release date, among others. However, it's essential to remember that IMDb scores are ultimately a reflection of audience opinions, and these opinions can change over time. Therefore, any prediction model should be periodically updated and validated against new data.