Movie Success Prediction

Presented By :Anas Pathan

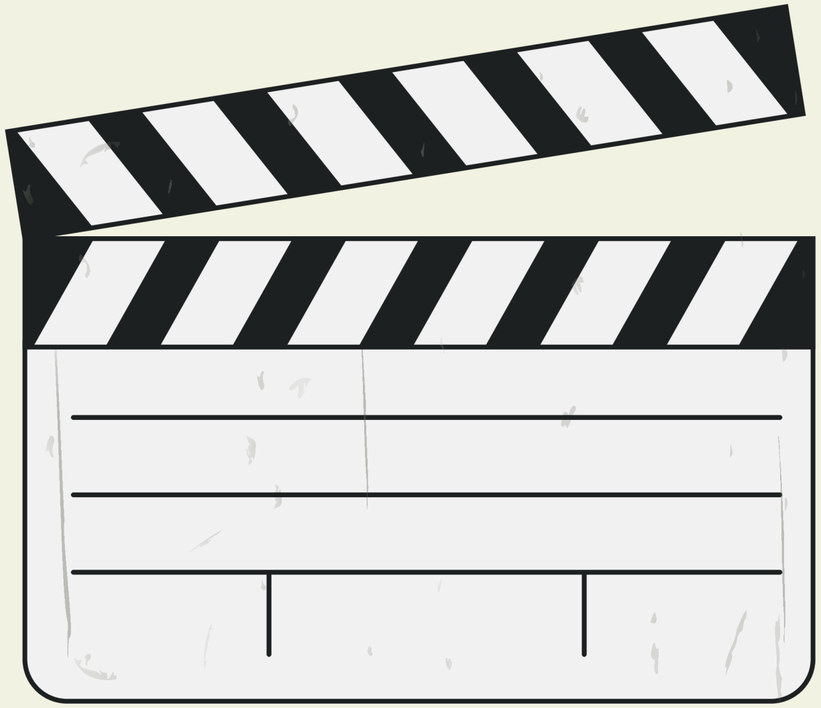
INTRODUCTION

- The **success of a movie** is influenced by a variety of factors, ranging from its budget to its genre, cast, and audience reception. Understanding these factors is crucial for producers, directors, and investors to make informed decisions and maximize the chances of a movie's success.

- With the power of **Data Science**, predictive models can analyze historical data and identify patterns that contribute to a *movie's success.* This allows stakeholders to estimate a movie's performance even before its release.

- This project focuses on a dataset that explores various attributes of movies, providing insights into the key drivers of success. The **dataset includes 28 significant features** such as a Budget, Gross, Duration and more.

- The target variable, '**IMDb Score,**' reflects the movie's success and popularity, offering a comprehensive measure to evaluate performance..



Objective OF The Study

Provide Insights for Decision-Making

Comparing Algorithms

Real - time Monitoring

Understanding Industry Trends

Predictive Insights for Better Outcomes

Enhanced Strategic Planning

VALUE OF THE STUDY

WORK FLOW

Data Collection

Gathering and organizing data to train machine learning models.

STEP 01

STEP 02

Exploratory Data Analysis

Analyzing and visualizing data patterns to understand its characteristics

PREPROCESSING

Preparing and cleaning data to enhance its quality and suitability

STEP 03

STEP 04

Split Train And Test Data

Dividing the dataset into training and testing sets to evaluate

Model Selection and Model Training

Choosing a suitable machine learning algorithm and optimizing its parameters

STEP 05

STEP 06

Model Evaluation

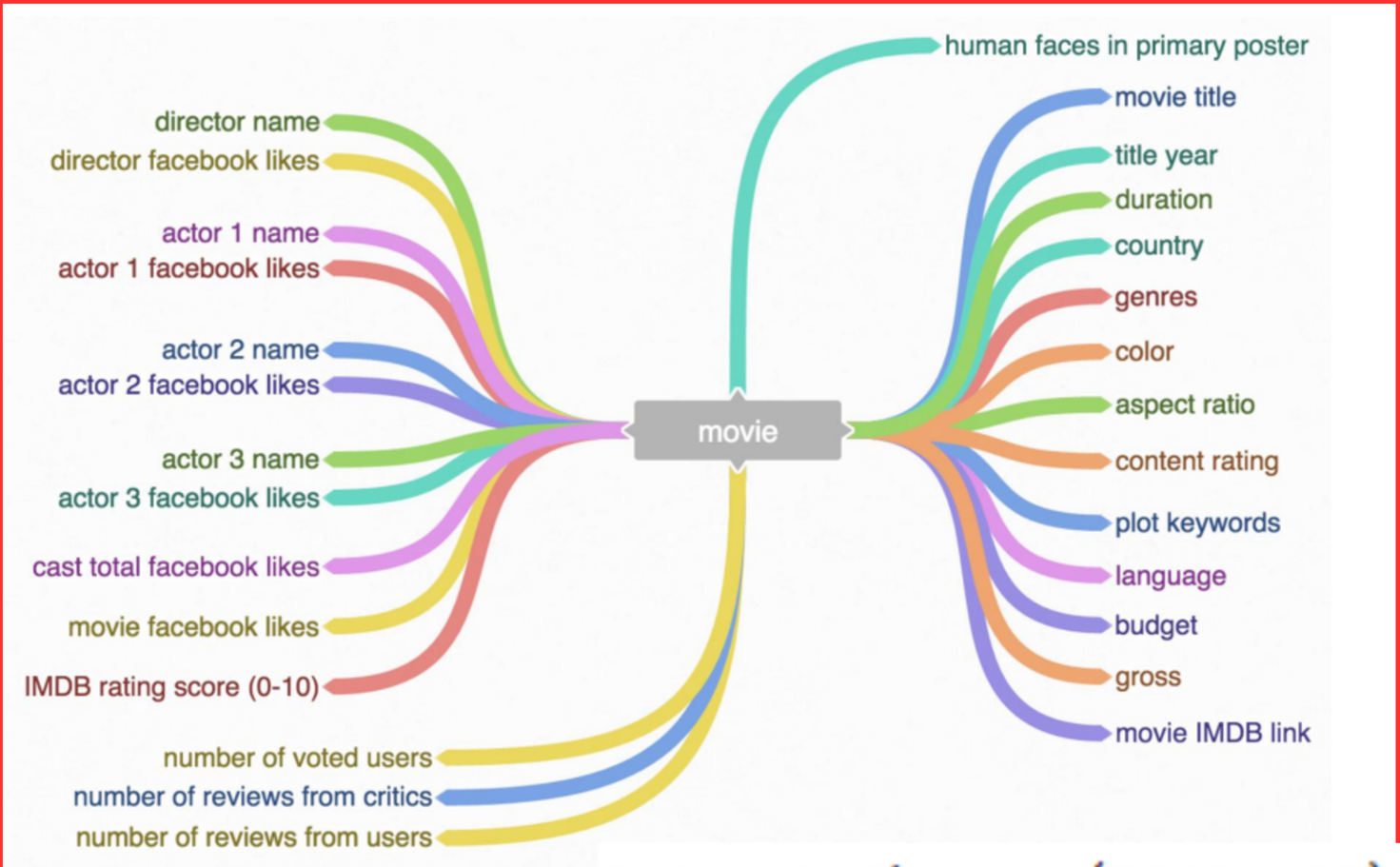
Assessing the performance of a machine learning model using metrics

Monitor and Update

Continuously monitor the model's performance in the real-world scenario

STEP 07

Data Gathering /Data Refinement

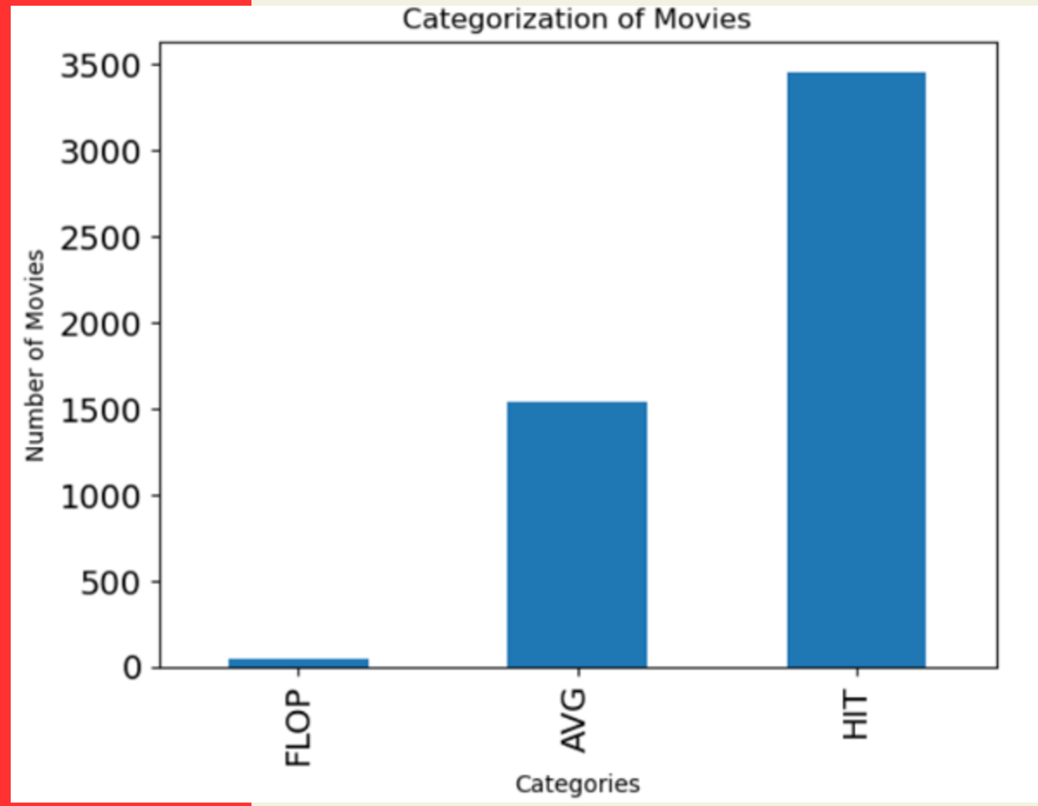


Dataset Shape: (5043, 28)

color
director_name
num_critic_for_reviews
duration
director_facebook_likes
actor_3_facebook_likes
actor_2_name
actor_1_facebook_likes
gross
genres
actor_1_name
movie_title
num_voted_users
cast_total_facebook_likes
actor_3_name
facenumber_in_poster
plot_keywords
movie_imdb_link
num_user_for_reviews
language
country
content_rating
budget
title_year
actor_2_facebook_likes
imdb_score
aspect_ratio
movie_facebook_likes

Categorizing the Target Variables:

- Creating a new column Classify to categorize movies into "Hit", "Average", or "Flop" based on the IMDB score ranges(|1-3 | -Flop Movie,|3-6 |- Average Movie,|6-10 |- Hit Movie)
- As seen in the graph there are more number of hit movies.



Handling Missing Values.

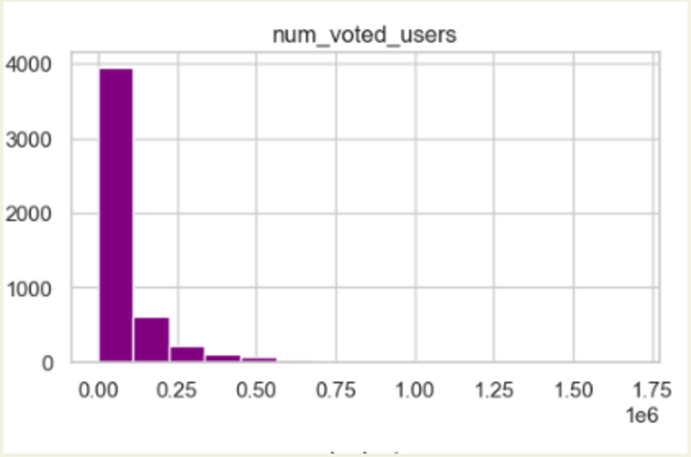
- Dropping the samples which have missing values.

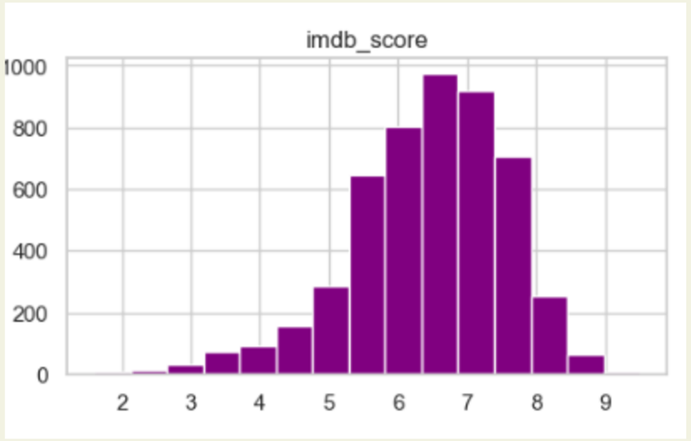
- After dropping all the samples which have missing values we are left with a clean data which has 3755 rows and 29 columns.
- No column has been dropped.

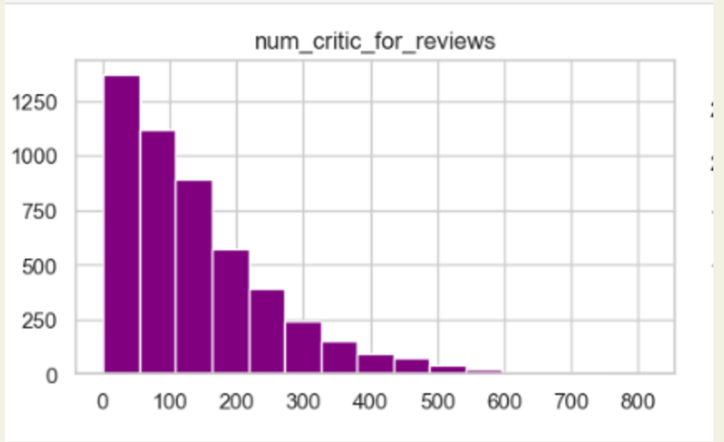
We save the clean data as a separate csv file for making a dashboard.

EXPLORATORY DATA ANALYSIS

- **Exploratory Data Analysis (EDA)** helped us understand the structure of the dataset, uncover patterns, identify trends, and gain valuable insights into the factors that contribute to a movie's success..
- Through EDA, we analyzed the distribution of various features and examined correlations between them to determine how they impact the target variable, IMDb Score .
- Using visualizations enabled us to clearly interpret the data, understand relationships between features like budget, duration, and genre, and pinpoint the key factors that play a vital role in predicting a movie's success.
- The dataset has been cleaned to handle NULL VALUES, but we identified some DUPLICATE RECORDS and OUTLIERS that required further handling to ensure the accuracy of our predictions.
- Additionally, we reviewed the data columns and found that certain numerical columns (e.g., content rating or genres represented as numbers) are categorical by semantics. We converted these columns into categorical (object) types for better analysis and interpretation

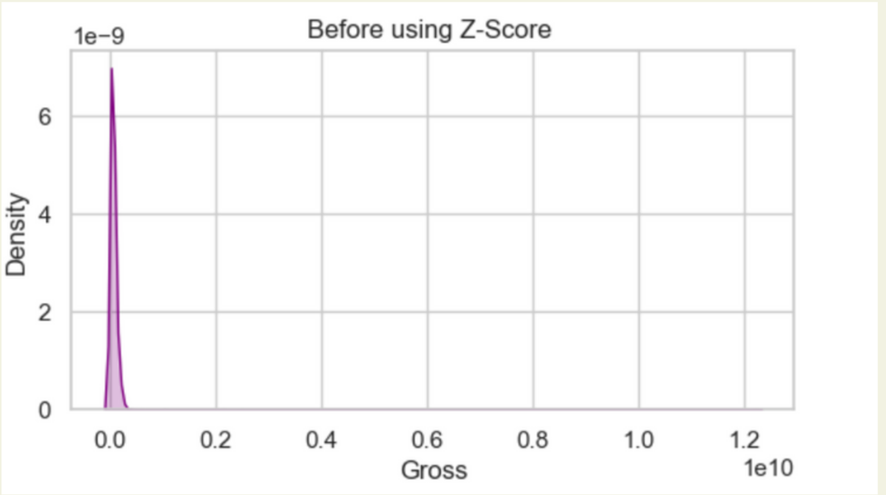


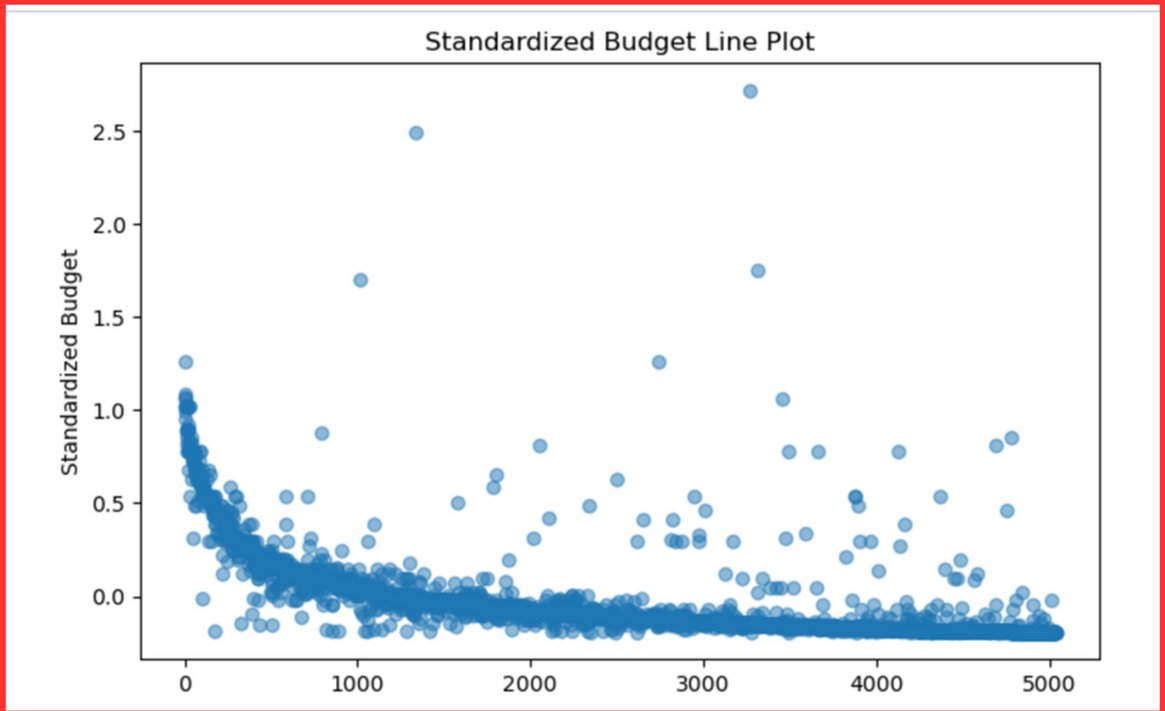


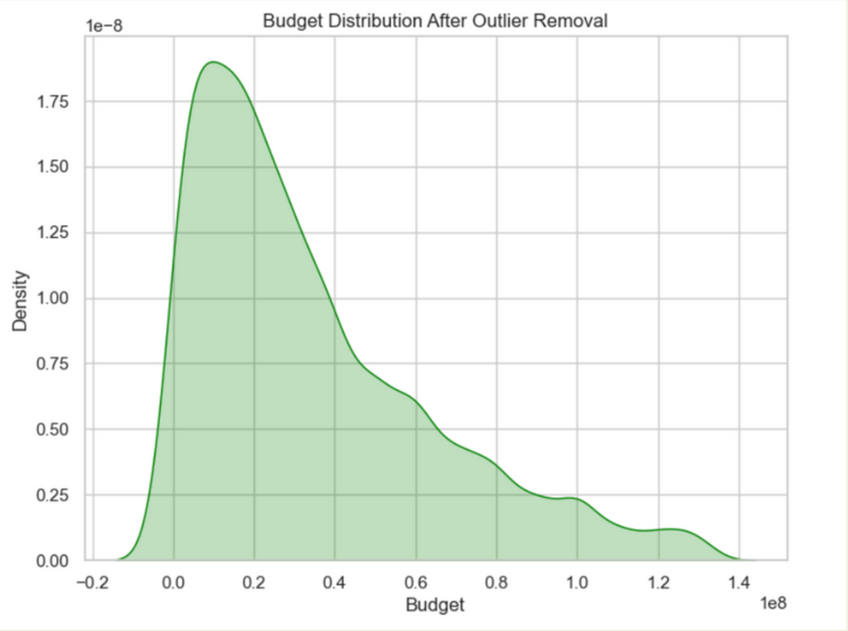


Outlier Detection

- Oultlier Detection Is most IMP part for model
- Mainly outliers Present in Budget and Gross Columns
- Used The Log F, Z Score & IQR For Removing Outliers







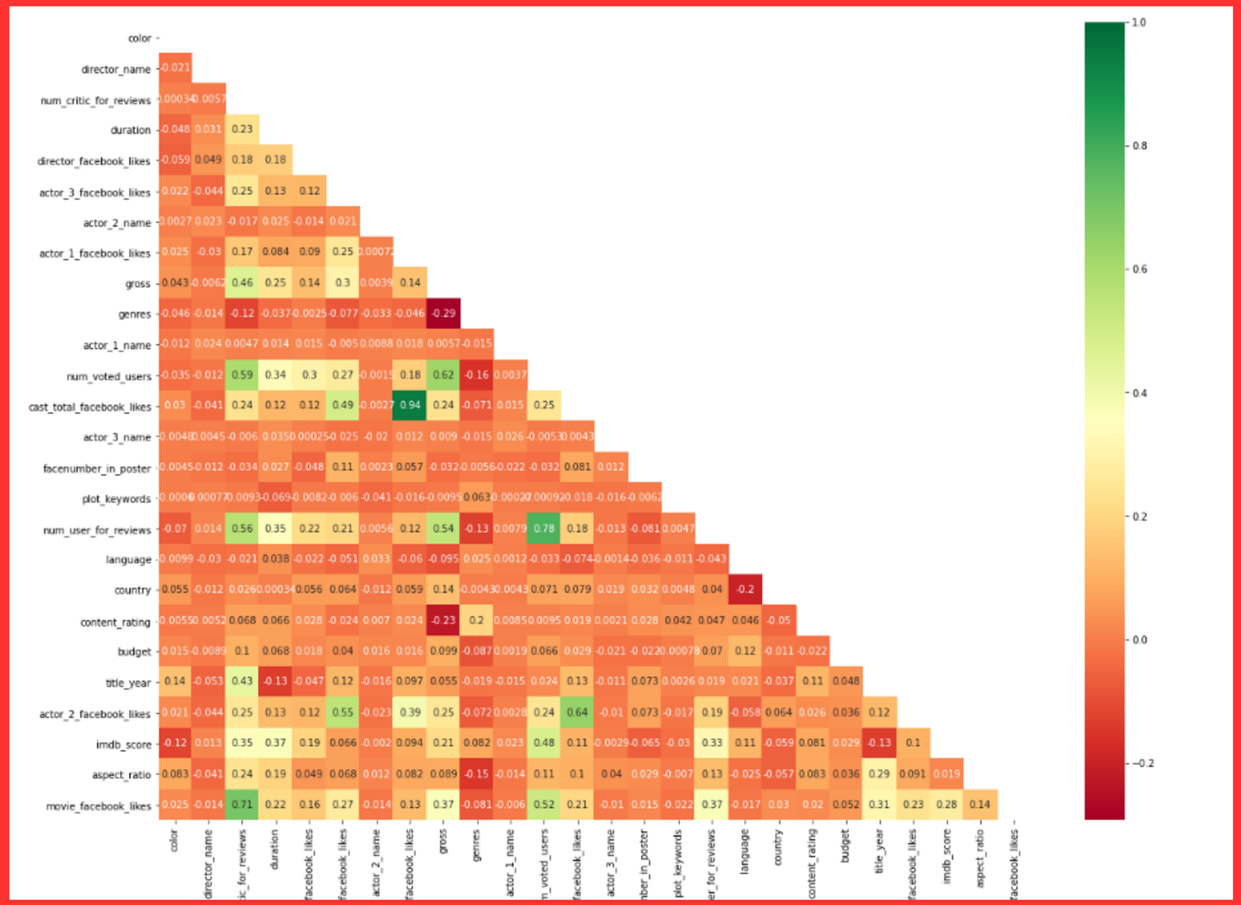
CORRELATION HEATMAP

Strong Positive Correlations

- num_voted_users ↔ imdb_score (around 0.7): Indicates that the number of user votes is strongly correlated with the IMDb score, which could be a significant feature for predicting movie success.

Weak or Negligible Correlations:

- content_rating ↔ Other Variables (near0)Content rating has little to no correlation with most variables, implying it might not significantly impact movie success.



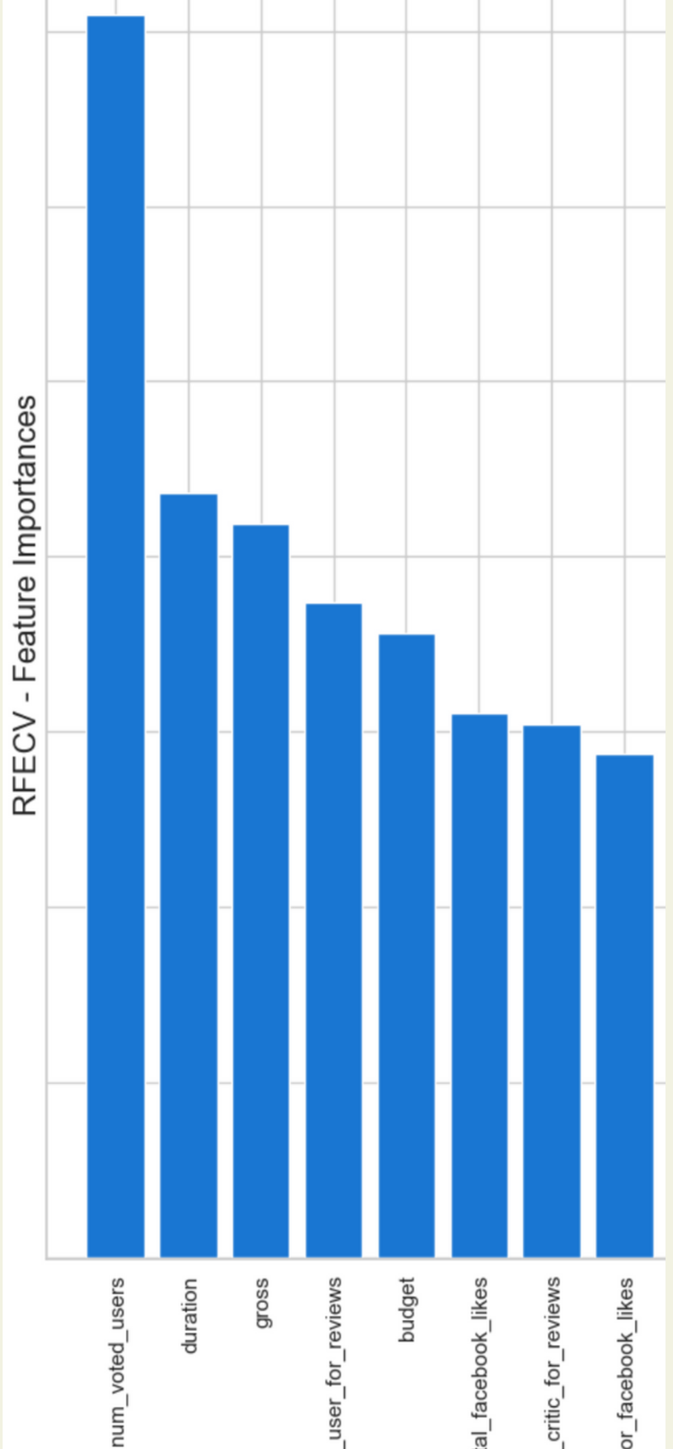
TRAIN TEST SPLIT

- We divided the data into training (70%) and testing (30%) sets.
- Setting a random state ensures **consistent results and** using stratify=**y** maintains a proportional **distribution** of the **target variable** in both sets.

SPLITING THE DATA INTO X & Y

- We divided the dataset into two parts: X and y.
- "**X**" typically represents the **independent** Variables, and "**y**" represents the **Dependent** (**target variable**) that we want to predict or understand.(IMDB Column)

Feature Selection:-



MODEL SELECTION

- Models used:

- **Logistic Regression:** logistic Regression is commonly used for **binary classification problems**. it's preferred because it provides a **simple** an **efficient** way to **model** the **relationship** between the **independent variables** and the **probability** of a certain **outcome**.

- **Decision Tree:** Decision Tree algorithms are used for **classification** because they are **simple**, **computationally efficient**, and **effective** in handling **high-dimensional data**.
  Works best for categorical independent columns.

- **Random Forest Algorithm:** Random Forest: Random Forest is a robust supervised algorithm suitable for both regression and classification tasks.

Result With Random Forest

```
Confusion Matrix for Random Forest:
 [[ 917    0   60]
 [   1   16    5]
 [  28    0 2107]]
Classification Report for Random Forest:
              precision    recall  f1-score   support

           0       0.97      0.94      0.95       977
           1       1.00      0.73      0.84        22
           2       0.97      0.99      0.98      2135

    accuracy                           0.97      3134
   macro avg       0.98      0.88      0.92      3134
weighted avg       0.97      0.97      0.97      3134

Random Forest Accuracy: 0.97
```

Result With Decision Tree

```
Confusion Matrix for Decision Tree:
 [[ 12   3]
 [  7 995]]
Classification Report for Decision Tree:
              precision    recall  f1-score   support

           0       0.63      0.80      0.71        15
           1       1.00      0.99      0.99      1002

    accuracy                           0.99      1017
   macro avg       0.81      0.90      0.85      1017
weighted avg       0.99      0.99      0.99      1017

Decision Tree Accuracy: 0.99
```

Result With Logistic

```
Confusion Matrix for Logistic Regression:
 [[   0   15]
 [   0 1002]]
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        15
           1       0.99      1.00      0.99      1002

    accuracy                           0.99      1017
   macro avg       0.49      0.50      0.50      1017
weighted avg       0.97      0.99      0.98      1017

Logistic Regression Accuracy: 0.99
```

Conclusion

- *The project successfully demonstrated the use of machine learning models to predict movie success. Both Decision Tree and Logistic Regression achieved high accuracy of 99%, indicating the effectiveness of the features selected and the data preprocessing steps.*
- Among the models, the Decision Tree was chosen for its interpretability and ability to handle complex decision-making processes, making it a robust choice for *predicting movie success.*
- *This study highlights the potential of data-driven approaches in the entertainment industry for making informed decisions.*

Thank You

BIA BOSTON INSTITUTE OF ANALYTICS ®