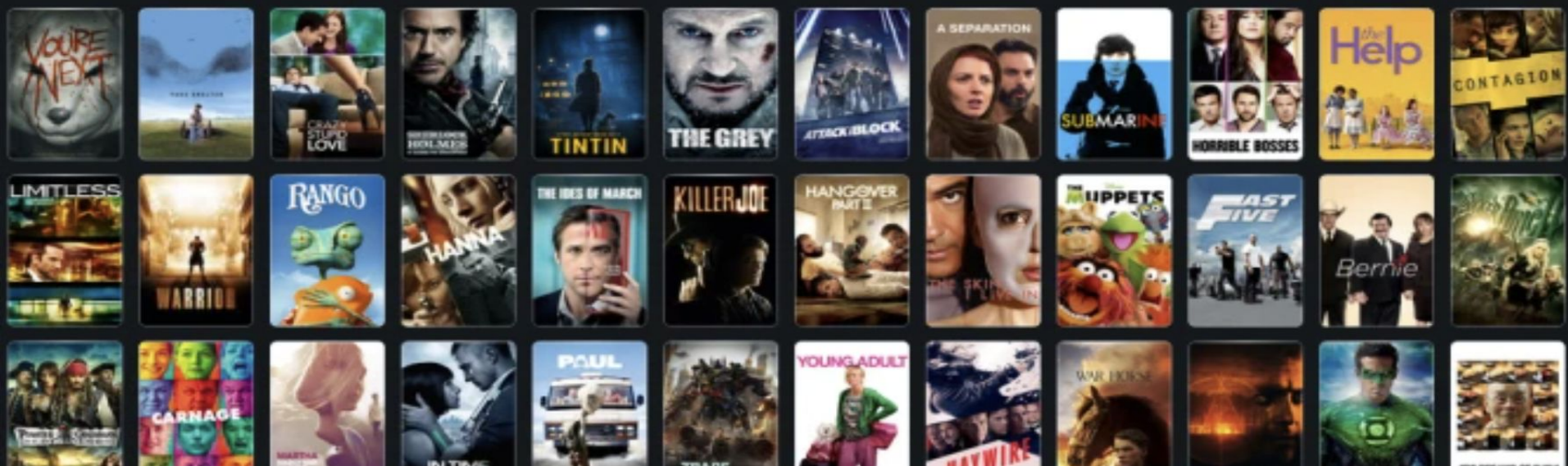


Predicting **IMDb** Movie Ratings



Group Members

Adegbenga Ayoola

Grace Le

Michelle Raj

Syed Bari

Wardah Anis

Objective

The purpose of the project is to predict the IMDb Movie Rating from the features contained in the IMDb dataset, execute exploratory Data Analysis and Visualization for the IMDb Dataset

And finally utilise Machine Learning Models to achieve ideal model performance by comparing suitable Machine Learning Algorithms

Why we chose the topic?

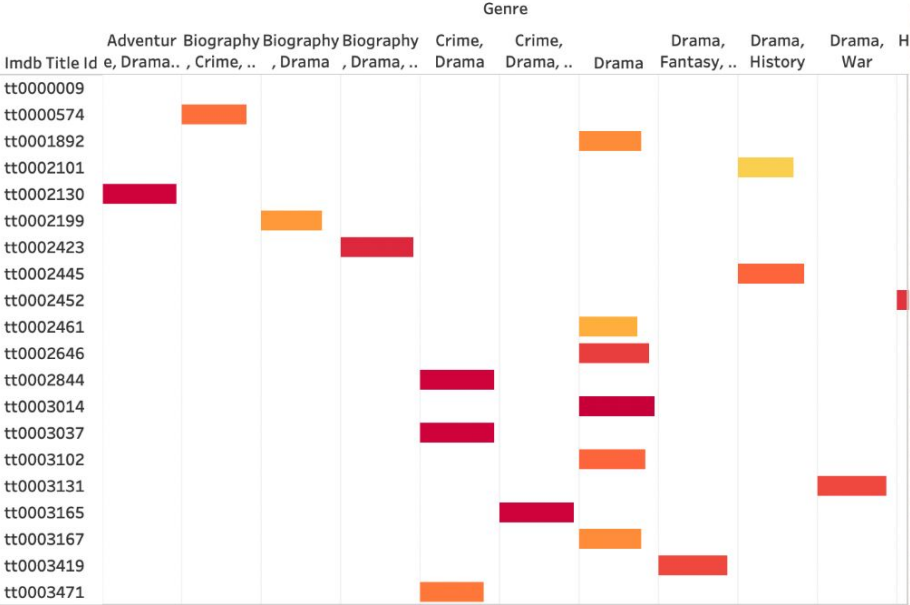
- IMDb, whose full English name is Internet Movie Database, is a web-based data set identified with films, TV shows, home recordings, games... Also, streaming internet-based substance, including full team: entertainers, creation group. What's more, IMDb gives a wide scope of film-related data, be it individual memoirs, plot synopses, tests, or client audits. From that point, you can likewise effectively see the sorts of rankings dependent on a wide range of rules, so new clients can undoubtedly see the issues and content that get the most client consideration.
- We chose this topic because it makes it easy for people to find good movies. Besides, with the current situation, the Covid-19 makes people still afraid to watch movies in cinemas, people often choose the movies to watch at home. Therefore, IMDb makes it easy for people to pick out which movies they love, with high ratings, or their favorite actors.

ERD Diagram

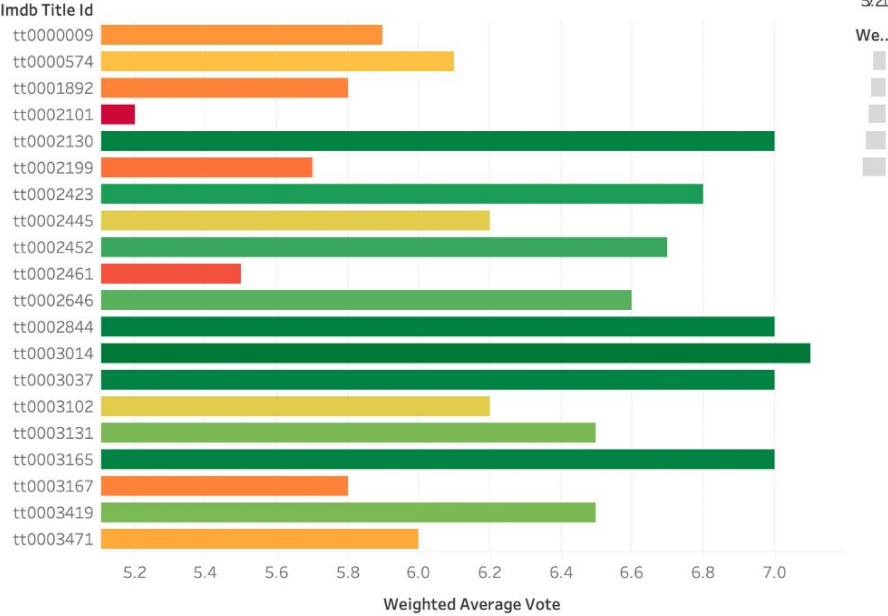


Dashboard

Genre vs Rating



Imdb ID vs Weighted Average Vote



Machine Learning Model Proposal








The purpose of the project is to find the best Machine learning Algorithms that can predict the imdb rating based on the features given in the four tables taken from our IMDB Dataset. To achieve this from Various Models the model with the lowest root mean squared error, best accuracy, and best confusion matrix is selected.

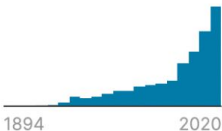
IMDB Movie Dataset Tables

Datasets selected from Kaggle

IMDB Movies.csv

△ imdb_title_id	# weighted_averag...	# total_votes	# mean_vote	# median_vote	# votes_10
title ID on IMDb	total weighted average rating	total votes received	total mean vote	total median vote	number of votes with rating equal to 10
85855 unique values					

IMDB Ratings.csv

△ imdb_title_id	△ title	△ original_title	# year	△ date_published	△ genre
title ID on IMDb	title name	original title name	year of release	date of release	movie genre
85855 unique values	82094 unique values	80852 unique values		22012 unique values	Drama 15% Comedy 9% Other (65619) 76%

Machine Learning Models

Linear Regression, Logistic Regression, Random Forest, SVM, K-means Algorithm

Target: IMDb Total Average Weighted Rating

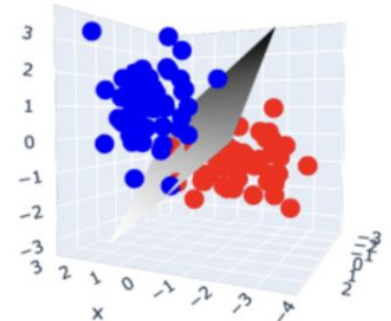
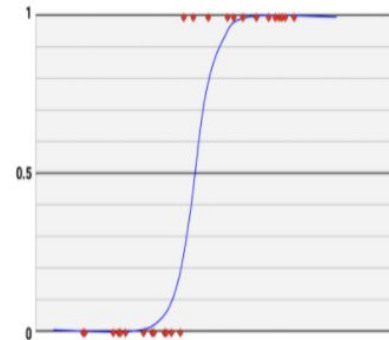
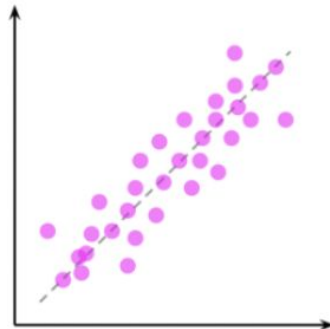
Output: IMDb Rating

Feature: imdb_title_id, title, year, date_published, genre, duration, country, language, director, writer, reviews

Categorical Feature: title, year, genre, country, language, direction, reviews

Quantitative Feature: Date, Duration, vote, budget, gross income, total votes, us_voters rating,

Results: Accuracy, Confusion Matrix



Multiple Linear Regression

Inputs, X: title, date_published, genre, duration, country, language, director, writer, reviews, title, year, genre, country, language, direction, reviews

Output, Y: Average Weighted Rating

Decision Tree:

Inputs, X: title, date_published, genre, duration, country, language, director, writer, reviews, title, year, genre, country, language, direction, reviews

Output, Y: Average Weighted Rating

Random Forest

Inputs, X: title, date_published, genre, duration, country, language, director, writer, reviews, title, year, genre, country, language, direction, reviews

Output, Y: Average Weighted Rating

SVM

Inputs, X: title, date_published, genre, duration, country, language, director, writer, reviews, title, year, genre, country, language, direction, reviews

Output, Y: Average Weighted Rating

Datatype of Inputs and Outputs

Inputs

title	object
original_title	object
year	object
date_published	object
genre	object
duration	int64
country	object
language	object
director	object
writer	object
production_company	object
actors	object
description	object
avg_vote	float64
votes	int64
budget	object
usa_gross_income	object
worldwide_gross_income	object
metascore	float64
reviews_from_users	float64
reviews_from_critics	float64
dtype:	object

Outputs

weighted_average_vote	float64
total_votes	int64
...	...

Summary of Significant Steps

1. Join the four tables to create a single table, analyze the datasets
2. Data Cleaning, and dropping unwanted rows and columns
3. Find the relationship between the feature and the target, find the importance of each feature, drop the columns according to the importance ranking
4. Data splitting into training and testing sets
5. Train and Fit the Machine Learning Model using the processed and cleaned data
6. Calculated the balanced accuracy score along with the confusion matrix
7. Compare accuracy in different models ; SVM, Random Forest and Neural Networks
8. Add results to database such as Postgres pgAdmin
9. Reports outcomes in Tableau for Visualization.
10. Final Summary

Thank You