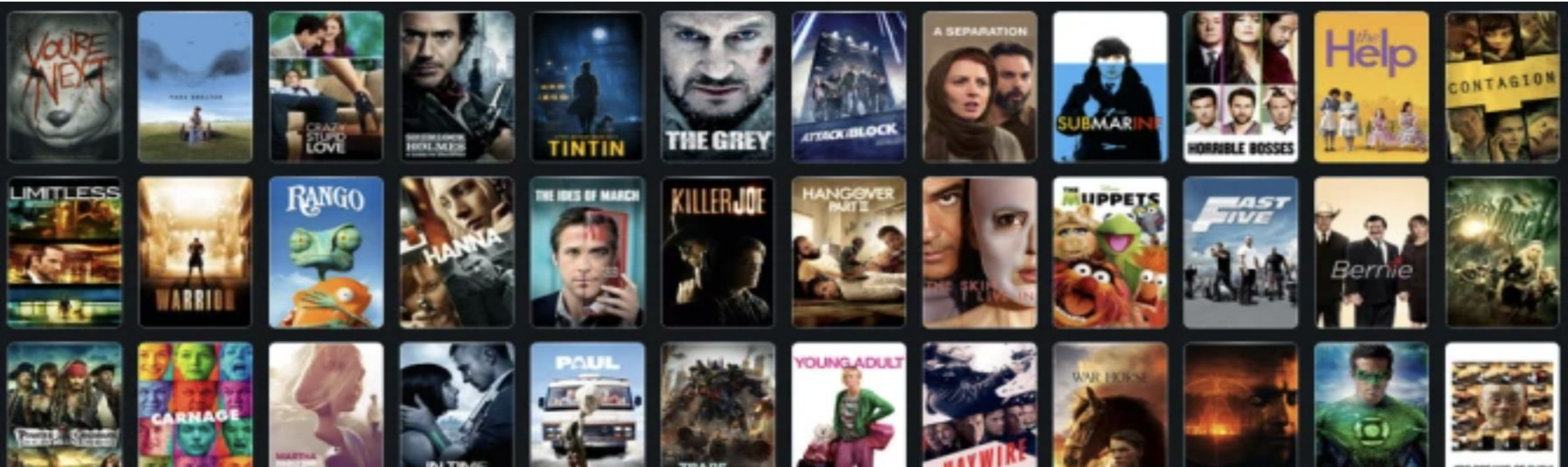
Predicting IMDb Movie Ratings



Group Members

Adegbenga Ayoola
Grace Le
Michelle Raj
Syed Bari
Wardah Anis

OBJECTIVE

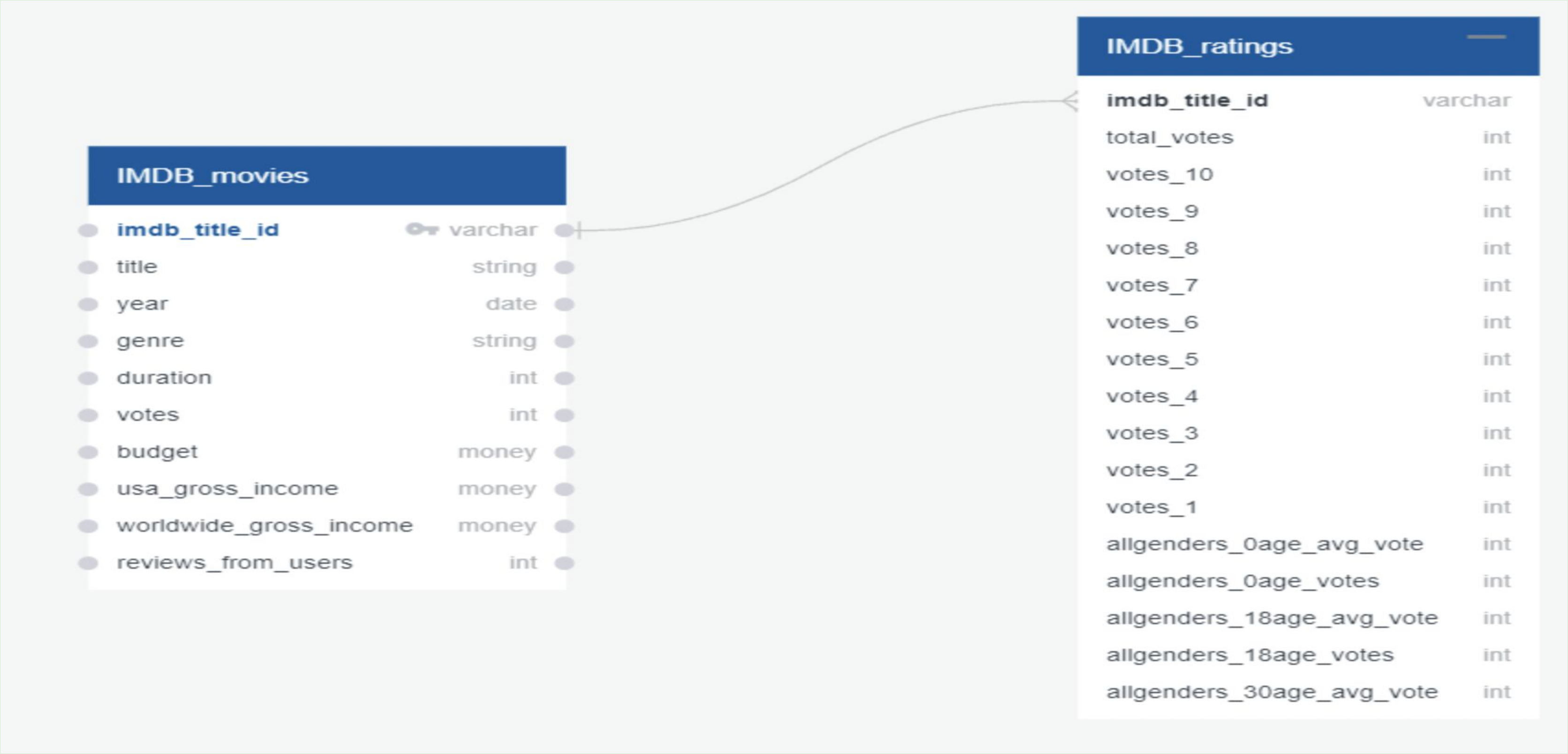
The purpose of the project is to predict the IMDb Movie Rating from the features contained in the IMDb dataset, execute exploratory Data Analysis and Visualization for the IMDb Dataset
And finally utilise Machine Learning Models to achieve ideal model performance by comparing suitable Machine Learning Algorithms

Why we chose this 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.

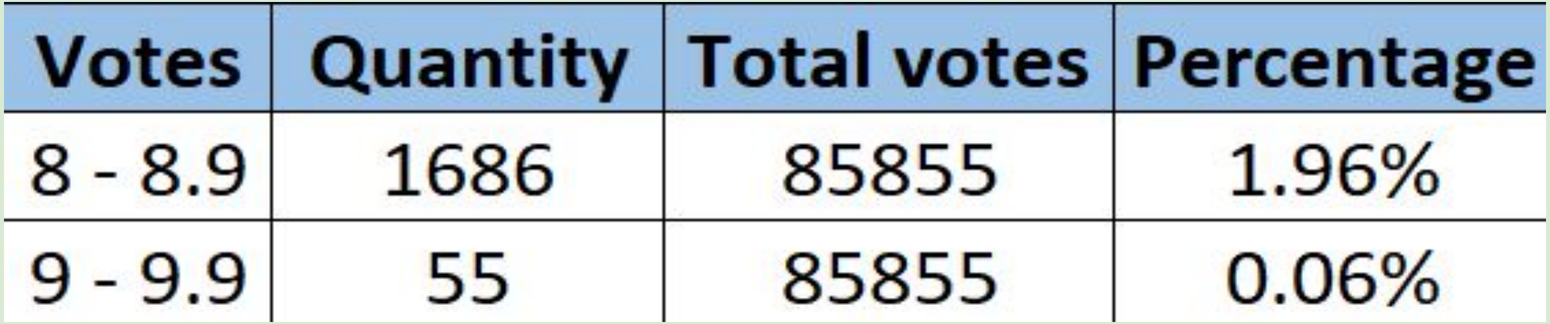
ERD Diagram



IMDB_movies	
imdb_title_id	varchar
title	string
year	date
genre	string
duration	int
votes	int
budget	money
usa_gross_income	money
worldwide_gross_income	money
reviews_from_users	int

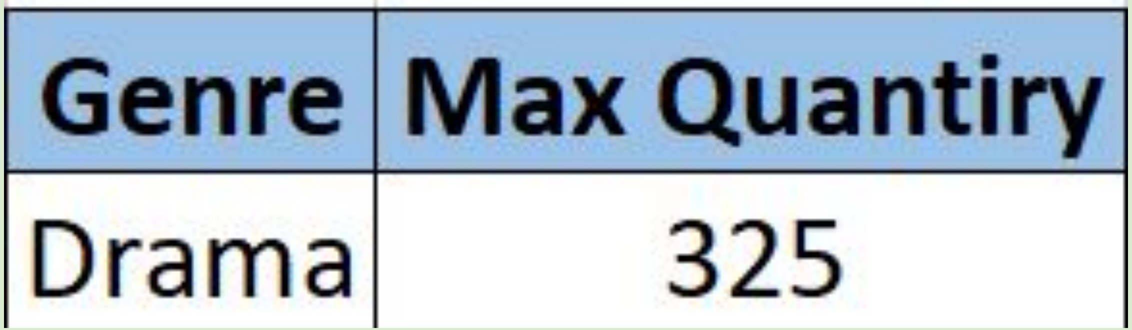
IMDB_ratings	
imdb_title_id	varchar
total_votes	int
votes_10	int
votes_9	int
votes_8	int
votes_7	int
votes_6	int
votes_5	int
votes_4	int
votes_3	int
votes_2	int
votes_1	int
allgenders_0age_avg_vote	int
allgenders_0age_votes	int
allgenders_18age_avg_vote	int
allgenders_18age_votes	int
allgenders_30age_avg_vote	int

Percentage of voting



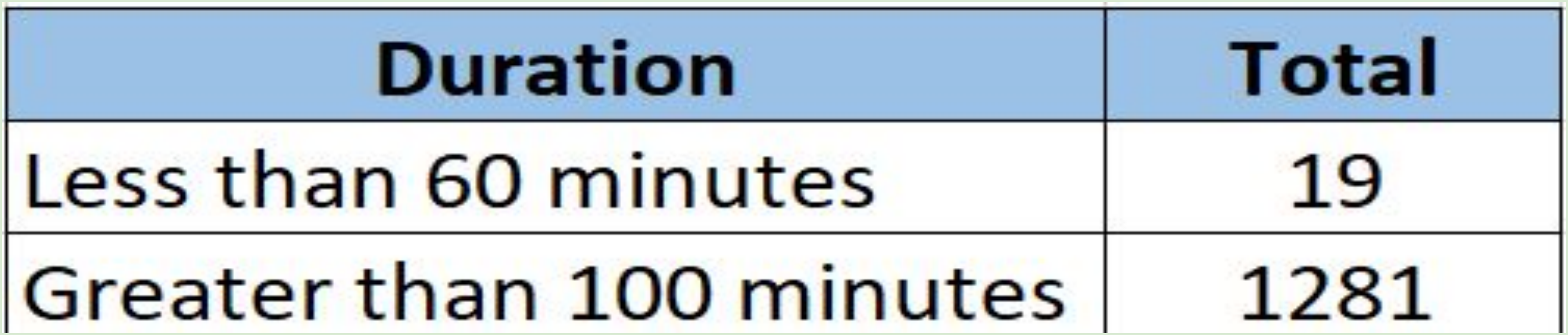
Votes	Quantity	Total votes	Percentage
8 - 8.9	1686	85855	1.96%
9 - 9.9	55	85855	0.06%

Highest genre of votes from 8 to 10



Genre	Max Quantiry
Drama	325

Comparing the duration from 8 to 10



Duration	Total
Less than 60 minutes	19
Greater than 100 minutes	1281

DataBase AWS S3

S3 is Amazon's cloud file storage service that uses key-value pairs. Files are stored on multiple servers and have a high rate of availability of more than 99.9%. S3 is also very scalable—you are not limited to the memory of one computer. As data flows in, more and more can be stored, as opposed to a local computer that is limited by available memory. Additionally, it offers availability—several team members can access massive amounts of data from one central location

AWS link for the IMDb datasets

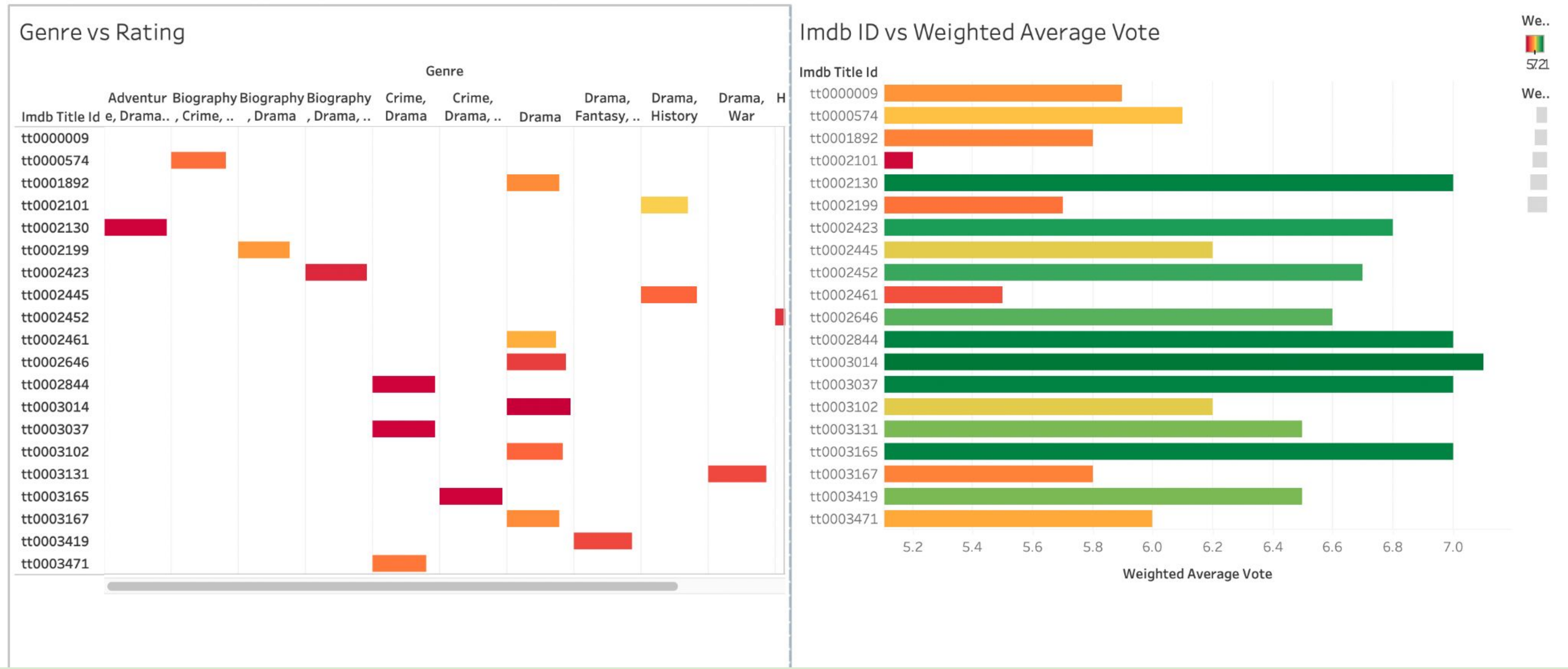
1. IMDb_ratings.csv Link

2. IMDb_movies.csv Link

3. Encoded_data.csv Link

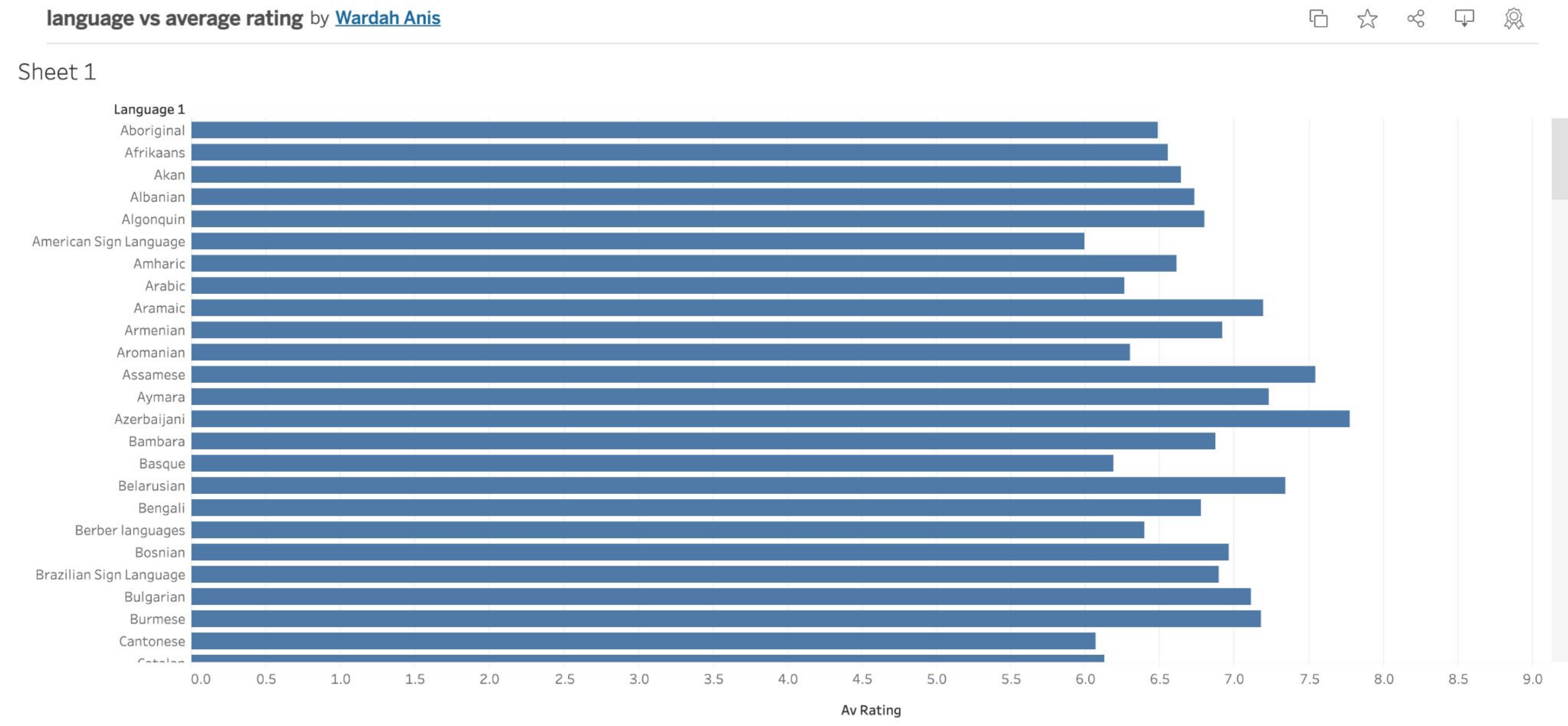
4. IMDb_data.csv Link

Dashboard



Genre vs Rating

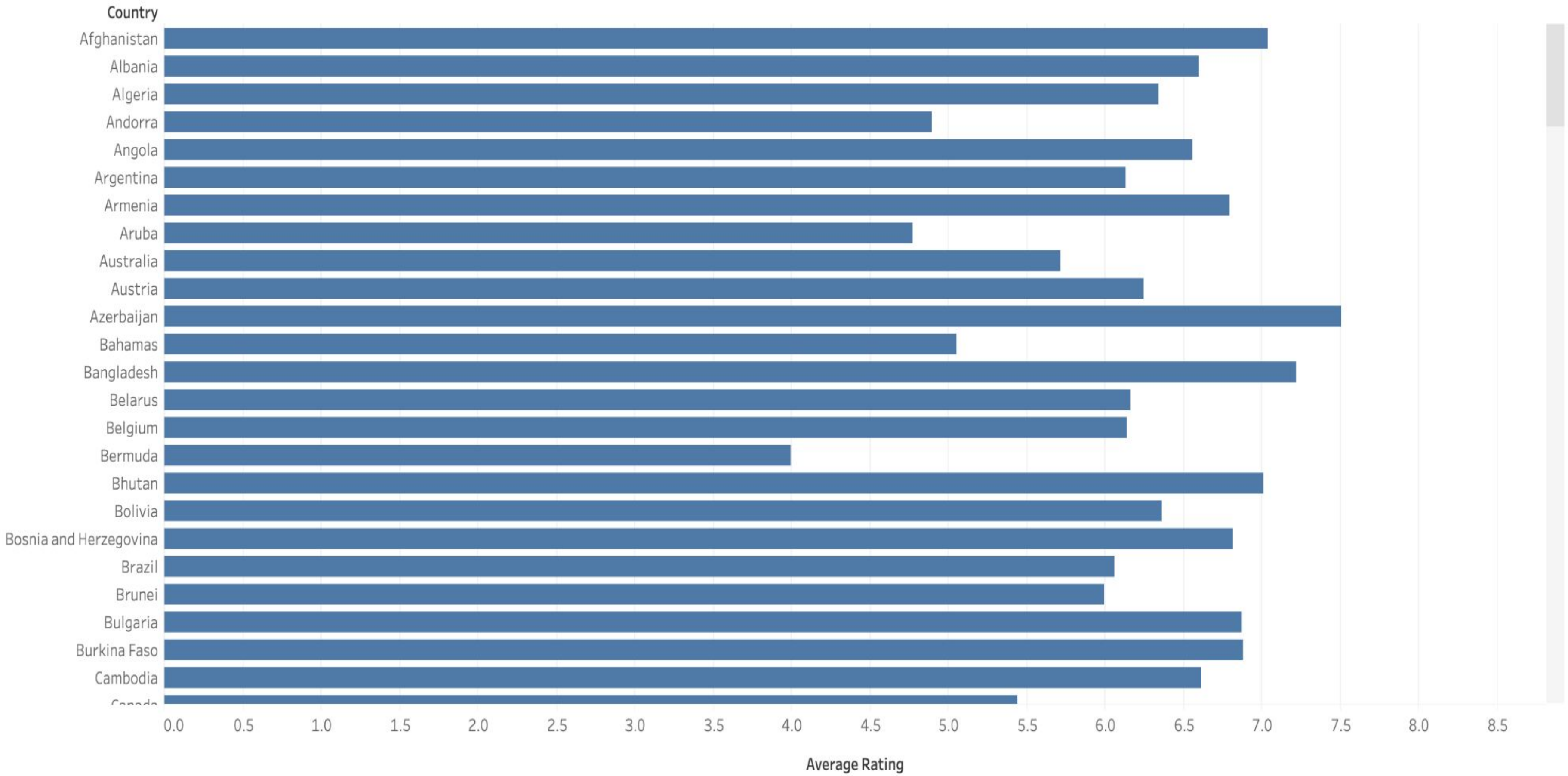
Imdb ID vs Weighted Average Vote

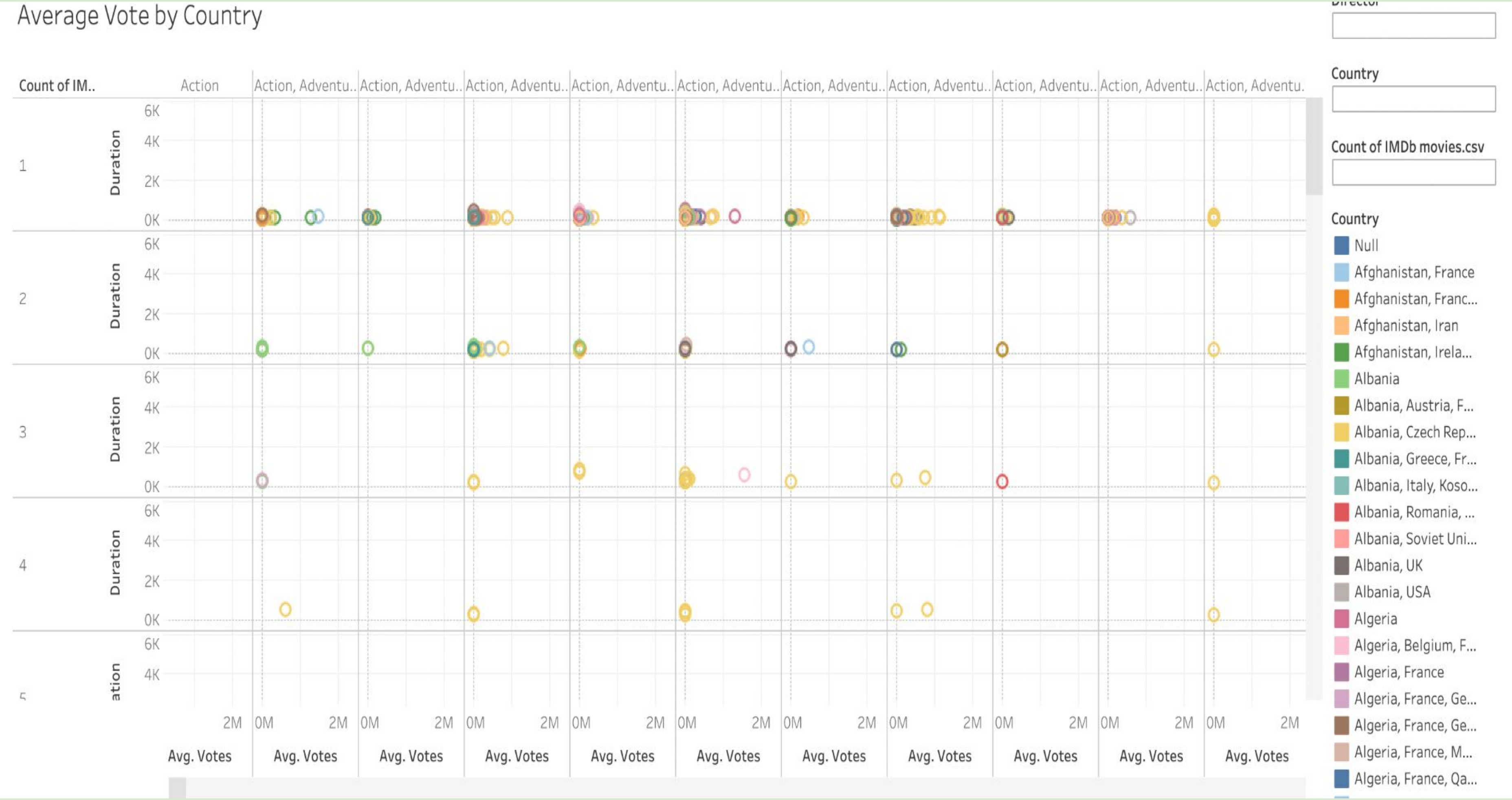


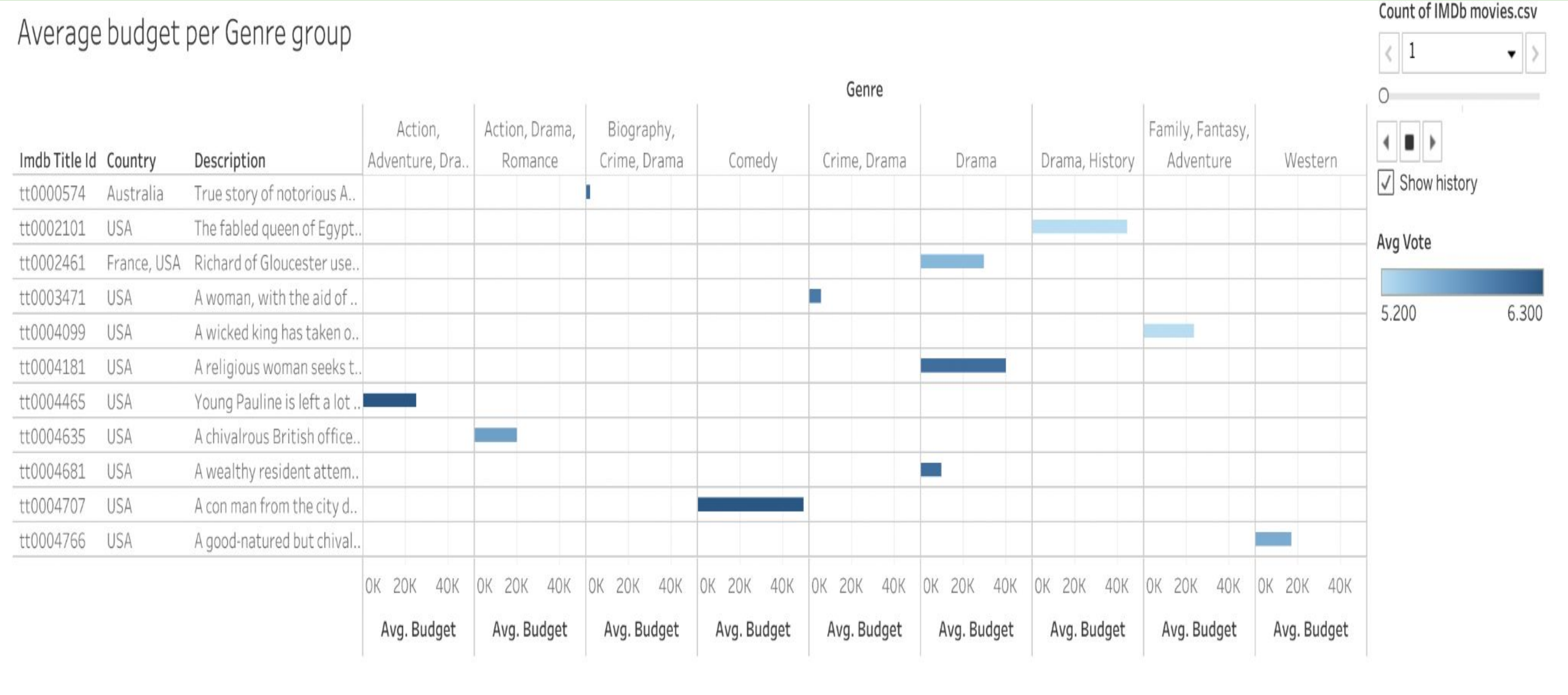
language vs average rating by Wardah Anis

Sheet 1

Country vs Average Rating




Average Vote by Country
Count of IM..
Action
Action, Adventu..
Action, Adventu..
Action, Adventu..
Action, Adventu..
Action, Adventu..
Action, Adventu..
Action, Adventu..
Action, Adventu..
Action, Adventu..
Action, Adventu..
1
2
3
4
5
Duration
Duration
Duration
Duration
ation
6K
4K
2K
0K
2M
0M
Avg. Votes
Director
Country
Count of IMDb movies.csv
Country
Null
Afghanistan, France
Afghanistan, Franc...
Afghanistan, Iran
Afghanistan, Irela...
Albania
Albania, Austria, F...
Albania, Czech Rep...
Albania, Greece, Fr...
Albania, Italy, Koso...
Albania, Romania, ...
Albania, Soviet Uni...
Albania, UK
Albania, USA
Algeria
Algeria, Belgium, F...
Algeria, France
Algeria, France, Ge...
Algeria, France, Ge...
Algeria, France, M...
Algeria, France, Qa...



Average budget per Genre group

Imdb Title Id	Country	Description
tt0000574	Australia	True story of notorious A..
tt0002101	USA	The fabled queen of Egypt..
tt0002461	France, USA	Richard of Gloucester use..
tt0003471	USA	A woman, with the aid of ..
tt0004099	USA	A wicked king has taken o..
tt0004181	USA	A religious woman seeks t..
tt0004465	USA	Young Pauline is left a lot ..
tt0004635	USA	A chivalrous British office..
tt0004681	USA	A wealthy resident attem..
tt0004707	USA	A con man from the city d..
tt0004766	USA	A good-natured but chival..

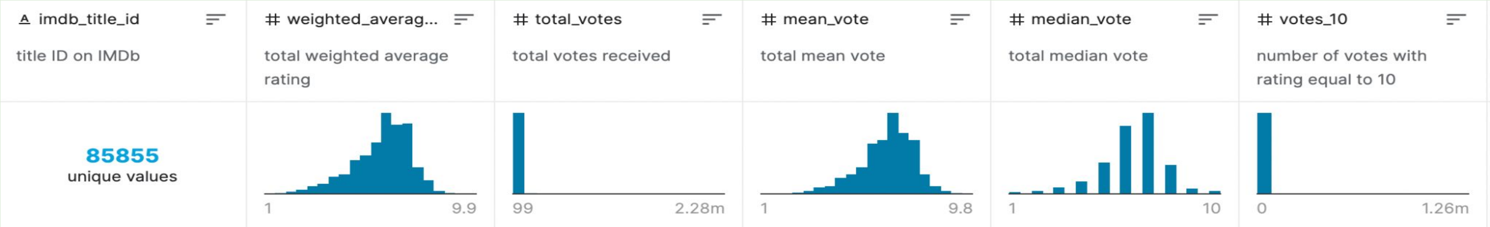
Machine Learning Model Proposal

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

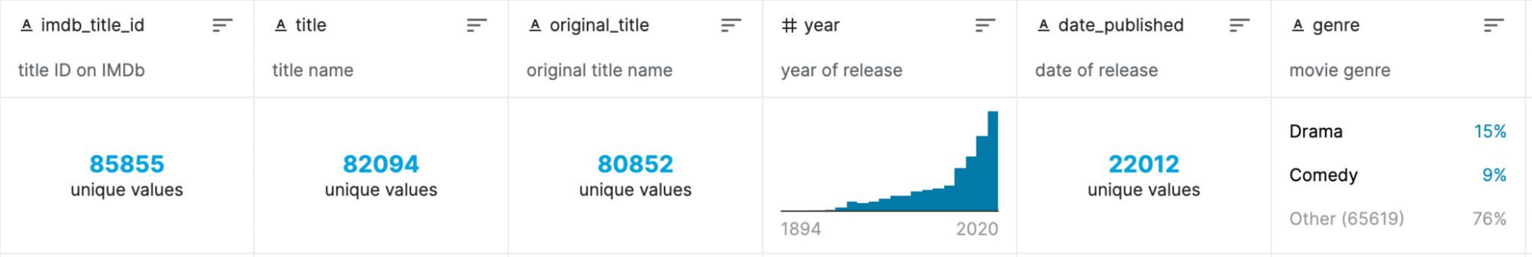
Datasets selected from Kaggle

IMDB Movies.csv



A 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	1 9.9	99 2.28m	1 9.8	1 10	0 1.26m

IMDB Ratings.csv



A imdb_title_id	A title	A original_title	# year	A date_published	A 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	1894 2020	22012 unique values	Drama 15% Comedy 9% Other (65619) 76%

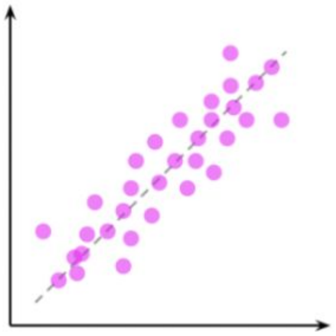
Machine Learning Models

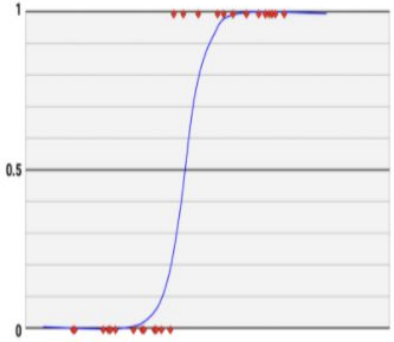
Linear Regression, Logistic Regression, Random Forest, SVM, Deep Learning, Gradient Boosting

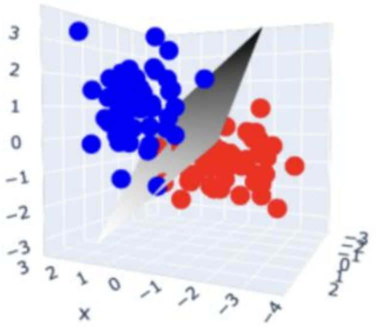
Target: IMDb Total Average Weighted Rating

Feature: genre, duration, country, language, weighted_average_rating, tot_voters_below_30, tot_voters_below_18, tot_voters_above_45, tot_voters_below_45, tot_male_voters, tot_female_voters
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, Mean Squared Error and Mean Absolute Error

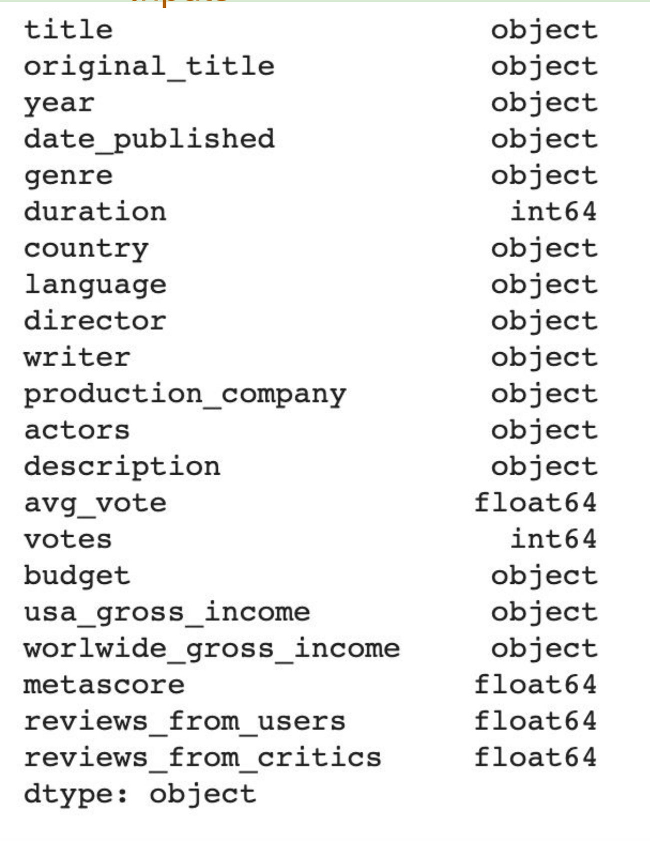






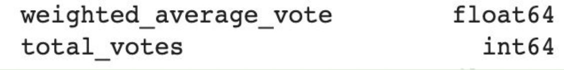
Datatype of Inputs and Outputs

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
worlwide_gross_income    object
metascore               float64
reviews_from_users      float64
reviews_from_critics    float64
dtype: object
```

Outputs



```
weighted_average_vote        float64
total_votes                    int64
```

Preliminary Data Preprocessing

- The very first steps include importing the libraries and importing the datasets which are IMDB_movies.csv a total of 22 columns and IMDB_ratings.csv a total of 49 columns

- The IMDB_movies.csv contained the features titel, year, genre, duration, country, language, director, actor, writer, description, average vote, budget, review from users, review from critics

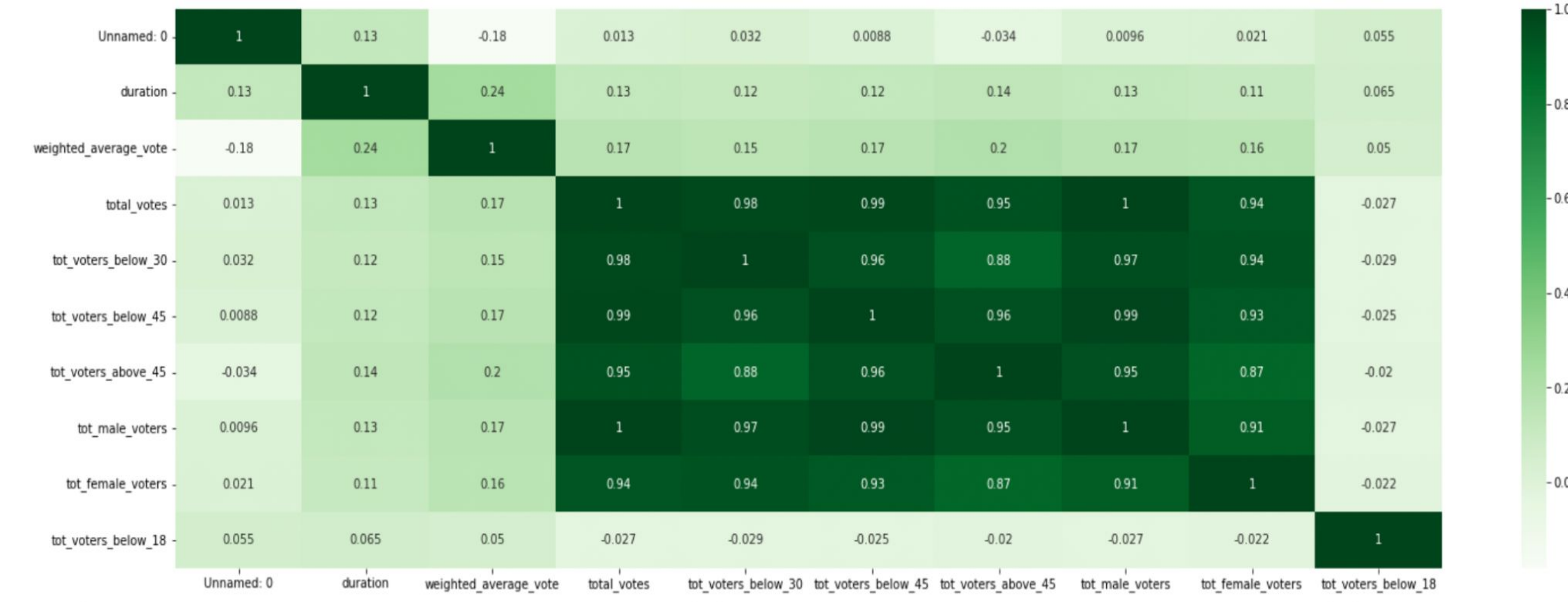
Description of preliminary feature engineering and selection

- Identify the dependent and the independent variable. After looking through the dataset the features removed were USA gross income, World gross income, metascore due to the null data. Other features were removed due to low correlation to the dependent variable "weighed average rating". The independent variable are as follows [genre, duration, country, language, weighted_average_rating, tot_voters_below_30, tot_voters_below_18, tot_voters_above_45, tot_voters_below_45, tot_male_voters, tot_female_voters]

- The dataset where the cleaned by addressing the null values and dropping hte null rows

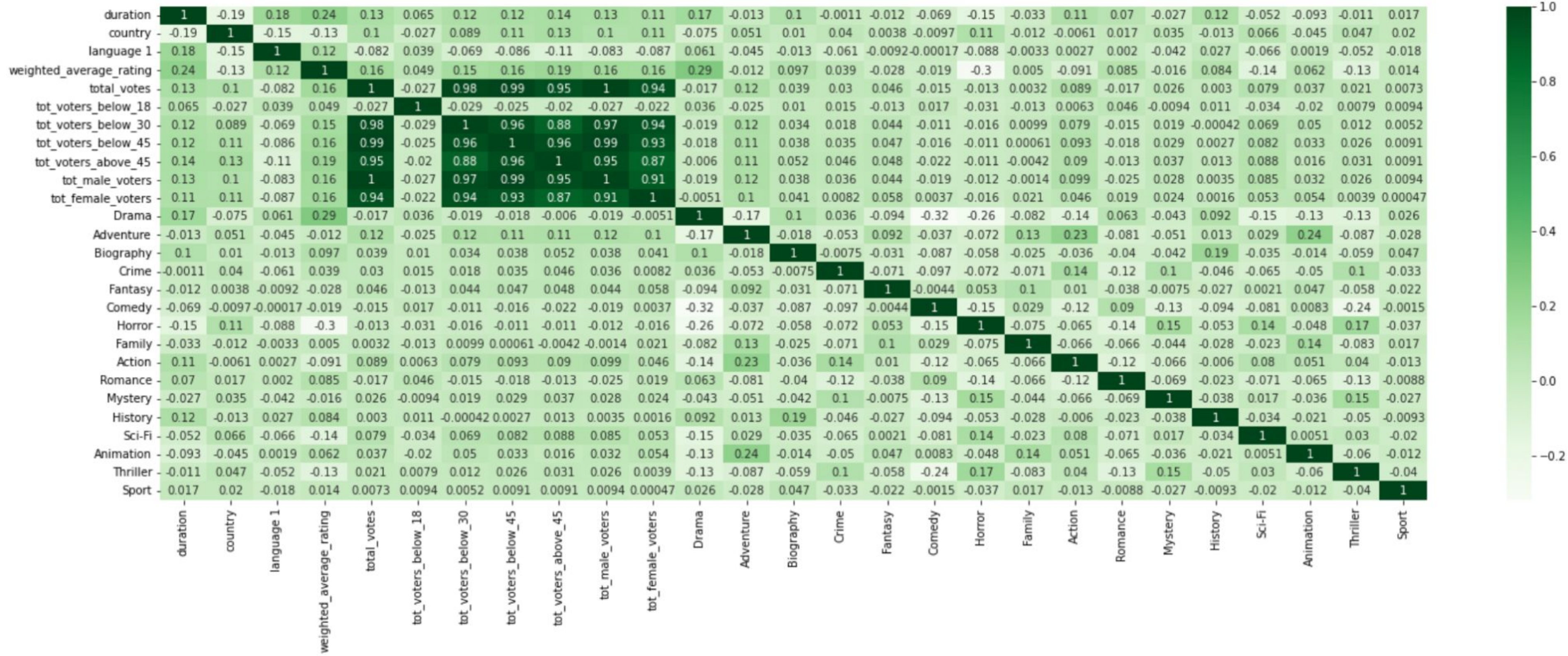
- After the cleaning process the categorical data like genre, country and language carrying multiple values were addressed using binary encoding with help of excel sheets

Correlation Matrix for Cleaned Data



	Unnamed: 0	duration	weighted_average_vote	total_votes	tot_voters_below_30	tot_voters_below_45	tot_voters_above_45	tot_male_voters	tot_female_voters	tot_voters_below_18
Unnamed: 0	1	0.13	-0.18	0.013	0.032	0.0088	-0.034	0.0096	0.021	0.055
duration	0.13	1	0.24	0.13	0.12	0.12	0.14	0.13	0.11	0.065
weighted_average_vote	-0.18	0.24	1	0.17	0.15	0.17	0.2	0.17	0.16	0.05
total_votes	0.013	0.13	0.17	1	0.98	0.99	0.95	1	0.94	-0.027
tot_voters_below_30	0.032	0.12	0.15	0.98	1	0.96	0.88	0.97	0.94	-0.029
tot_voters_below_45	0.0088	0.12	0.17	0.99	0.96	1	0.96	0.99	0.93	-0.025
tot_voters_above_45	-0.034	0.14	0.2	0.95	0.88	0.96	1	0.95	0.87	-0.02
tot_male_voters	0.0096	0.13	0.17	1	0.97	0.99	0.95	1	0.91	-0.027
tot_female_voters	0.021	0.11	0.16	0.94	0.94	0.93	0.87	0.91	1	-0.022
tot_voters_below_18	0.055	0.065	0.05	-0.027	-0.029	-0.025	-0.02	-0.027	-0.022	1

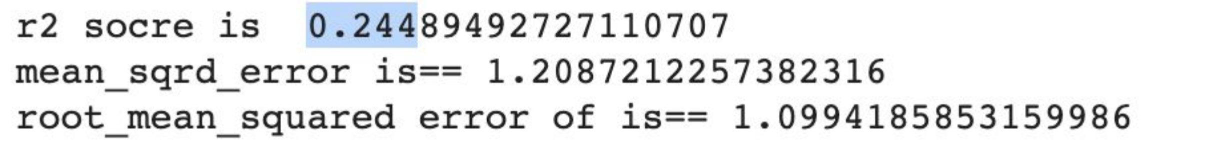
Correlation matrix of Encoded Data



	duration	country	language 1	weighted_average_rating	total_votes	tot_voters_below_18	tot_voters_below_30	tot_voters_below_45	tot_voters_above_45	tot_male_voters	tot_female_voters	Drama	Adventure	Biography	Crime	Fantasy	Comedy	Horror	Family	Action	Romance	Mystery	History	Sci-Fi	Animation	Thriller	Sport
duration	1	-0.19	0.18	0.24	0.13	0.065	0.12	0.12	0.14	0.13	0.11	0.17	-0.013	0.1	-0.0011	-0.012	-0.069	-0.15	-0.033	0.11	0.07	-0.027	0.12	-0.052	-0.093	-0.011	0.017
country	-0.19	1	-0.15	-0.13	0.1	-0.027	0.089	0.11	0.13	0.1	0.11	-0.075	0.051	0.01	0.04	0.0038	-0.0097	0.11	-0.012	-0.0061	0.017	0.035	-0.013	0.066	-0.045	0.047	0.02
language 1	0.18	-0.15	1	0.12	-0.082	0.039	-0.069	-0.086	-0.11	-0.083	-0.087	0.061	-0.045	-0.013	-0.061	-0.0092	-0.00017	-0.088	-0.0033	0.0027	0.002	-0.042	0.027	-0.066	0.0019	-0.052	-0.018
weighted_average_rating	0.24	-0.13	0.12	1	0.16	0.049	0.15	0.16	0.19	0.16	0.16	0.29	-0.012	0.097	0.039	-0.028	-0.019	-0.3	0.005	-0.091	0.085	-0.016	0.084	-0.14	0.062	-0.13	0.014
total_votes	0.13	0.1	-0.082	0.16	1	-0.027	0.98	0.99	0.95	1	0.94	-0.017	0.12	0.039	0.03	0.046	-0.015	-0.013	0.0032	0.089	-0.017	0.026	0.003	0.079	0.037	0.021	0.0073
tot_voters_below_18	0.065	-0.027	0.039	0.049	-0.027	1	-0.029	-0.025	-0.02	-0.027	-0.022	0.036	-0.025	0.01	0.015	-0.013	0.017	-0.031	-0.013	0.0063	0.046	-0.0094	0.011	-0.034	-0.02	0.0079	0.0094
tot_voters_below_30	0.12	0.089	-0.069	0.15	0.98	-0.029	1	0.96	0.88	0.97	0.94	-0.019	0.12	0.034	0.018	0.044	-0.011	-0.016	0.0099	0.079	-0.015	0.019	-0.00042	0.069	0.05	0.012	0.0052
tot_voters_below_45	0.12	0.11	-0.086	0.16	0.99	-0.025	0.96	1	0.96	0.99	0.93	-0.018	0.11	0.038	0.035	0.047	-0.016	-0.011	0.00061	0.093	-0.018	0.029	0.0027	0.082	0.033	0.026	0.0091
tot_voters_above_45	0.14	0.13	-0.11	0.19	0.95	-0.02	0.88	0.96	1	0.95	0.87	-0.006	0.11	0.052	0.046	0.048	-0.022	-0.011	-0.0042	0.09	-0.013	0.037	0.013	0.088	0.016	0.031	0.0091
tot_male_voters	0.13	0.1	-0.083	0.16	1	-0.027	0.97	0.99	0.95	1	0.91	-0.019	0.12	0.038	0.036	0.044	-0.019	-0.012	-0.0014	0.099	-0.025	0.028	0.0035	0.085	0.032	0.026	0.0094
tot_female_voters	0.11	0.11	-0.087	0.16	0.94	-0.022	0.94	0.93	0.87	0.91	1	-0.0051	0.1	0.041	0.0082	0.058	0.0037	-0.016	0.021	0.046	0.019	0.024	0.0016	0.053	0.054	0.0039	0.00047
Drama	0.17	-0.075	0.061	0.29	-0.017	0.036	-0.019	-0.018	-0.006	-0.019	-0.0051	1	-0.17	0.1	0.036	-0.094	-0.32	-0.26	-0.082	-0.14	0.063	-0.043	0.092	-0.15	-0.13	-0.13	0.026
Adventure	-0.013	0.051	-0.045	-0.012	0.12	-0.025	0.12	0.11	0.11	0.12	0.1	-0.17	1	-0.018	-0.053	0.092	-0.037	-0.072	0.13	0.23	-0.081	-0.051	0.013	0.029	0.24	-0.087	-0.028
Biography	0.1	0.01	-0.013	0.097	0.039	0.01	0.034	0.038	0.052	0.038	0.041	0.1	-0.018	1	-0.0075	-0.031	-0.087	-0.058	-0.025	-0.036	-0.04	-0.042	0.19	-0.035	-0.014	-0.059	0.047
Crime	-0.0011	0.04	-0.061	0.039	0.03	0.015	0.018	0.035	0.046	0.036	0.0082	0.036	-0.053	-0.0075	1	-0.071	-0.097	-0.072	-0.071	0.14	-0.12	0.1	-0.046	-0.065	-0.05	0.1	-0.033
Fantasy	-0.012	0.0038	-0.0092	-0.028	0.046	-0.013	0.044	0.047	0.048	0.044	0.058	-0.094	0.092	-0.031	-0.071	1	-0.0044	0.053	0.1	0.01	-0.038	-0.0075	-0.027	0.0021	0.047	-0.058	-0.022
Comedy	-0.069	-0.0097	-0.00017	-0.019	-0.015	0.017	-0.011	-0.016	-0.022	-0.019	0.0037	-0.32	-0.037	-0.087	-0.097	-0.0044	1	-0.15	0.029	-0.12	0.09	-0.13	-0.094	-0.081	0.0083	-0.24	-0.0015
Horror	-0.15	0.11	-0.088	-0.3	-0.013	-0.031	-0.016	-0.011	-0.011	-0.012	-0.016	-0.26	-0.072	-0.058	-0.072	0.053	-0.15	1	-0.075	-0.065	-0.14	0.15	-0.053	0.14	-0.048	0.17	-0.037
Family	-0.033	-0.012	-0.0033	0.005	0.0032	-0.013	0.0099	0.00061	-0.0042	-0.0014	0.021	-0.082	0.13	-0.025	-0.071	0.1	0.029	-0.075	1	-0.066	-0.066	-0.044	-0.028	-0.023	0.14	-0.083	0.017
Action	0.11	-0.0061	0.0027	-0.091	0.089	0.0063	0.079	0.093	0.09	0.099	0.046	-0.14	0.23	-0.036	0.14	0.01	-0.12	-0.065	-0.066	1	-0.12	-0.066	-0.006	0.08	0.051	0.04	-0.013
Romance	0.07	0.017	0.002	0.085	-0.017	0.046	-0.015	-0.018	-0.013	-0.025	0.019	0.063	-0.081	-0.04	-0.12	-0.038	0.09	-0.14	-0.066	-0.12	1	-0.069	-0.023	-0.071	-0.065	-0.13	-0.0088
Mystery	-0.027	0.035	-0.042	-0.016	0.026	-0.0094	0.019	0.029	0.037	0.028	0.024	-0.043	-0.051	-0.042	0.1	-0.0075	-0.13	0.15	-0.044	-0.066	-0.069	1	-0.038	0.017	-0.036	0.15	-0.027
History	0.12	-0.013	0.027	0.084	0.003	0.011	-0.00042	0.0027	0.013	0.0035	0.0016	0.092	0.013	0.19	-0.046	-0.027	-0.094	-0.053	-0.028	-0.006	-0.023	-0.038	1	-0.034	-0.021	-0.05	-0.0093
Sci-Fi	-0.052	0.066	-0.066	-0.14	0.079	-0.034	0.069	0.082	0.088	0.085	0.053	-0.15	0.029	-0.035	-0.065	0.0021	-0.081	0.14	-0.023	0.08	-0.071	0.017	-0.034	1	0.0051	0.03	-0.02
Animation	-0.093	-0.045	0.0019	0.062	0.037	-0.02	0.05	0.033	0.016	0.032	0.054	-0.13	0.24	-0.014	-0.05	0.047	0.0083	-0.048	0.14	0.051	-0.065	-0.036	-0.021	0.0051	1	-0.06	-0.012
Thriller	-0.011	0.047	-0.052	-0.13	0.021	0.0079	0.012	0.026	0.031	0.026	0.0039	-0.13	-0.087	-0.059	0.1	-0.058	-0.24	0.17	-0.083	0.04	-0.13	0.15	-0.05	0.03	-0.06	1	-0.04
Sport	0.017	0.02	-0.018	0.014	0.0073	0.0094	0.0052	0.0091	0.0091	0.0094	0.00047	0.026	-0.028	0.047	-0.033	-0.022	-0.0015	-0.037	0.017	-0.013	-0.0088	-0.027	-0.0093	-0.02	-0.012	-0.04	1

1. Multiple Linear Regression

Linear Regression is a very simple algorithm that can be implemented very easily to give satisfactory results.Furthermore, these models can be trained easily and efficiently even on systems with relatively low computational power when compared to other complex algorithms.Linear regression has a considerably lower time complexity when compared to some of the other machine learning algorithms.The mathematical equations of Linear regression are also fairly easy to understand and interpret



```
r2 socre is  0.24489492727110707
mean_sqrd_error is== 1.2087212257382316
root_mean_squared error of is== 1.0994185853159986
```

The reason for the low score would be due to insufficent data relating to the average rating variable and relevance to the problem although the score has increased from previous linear regression model by 20% more.

Benefits

- Models can be trained easily and efficiently even on systems with relatively low computational power when compared to other complex algorithms
- Linear regression fits linearly separable datasets almost perfectly and is often used to find the nature of the relationship between variables
- Overfitting is a situation that arises when a machine learning model fits a dataset very closely and hence captures the noisy data as well

Limitation

- A situation that arises when a machine learning model fails to capture the data properly.This typically occurs when the hypothesis function cannot fit the data well.
- **Outliers** of a data set are anomalies or extreme values that deviate from the other data points of the distribution.Data outliers can damage the performance
- Outliers can have a very big impact on linear regression's performance and hence they must be dealt with appropriately before linear regression is applied on the dataset.

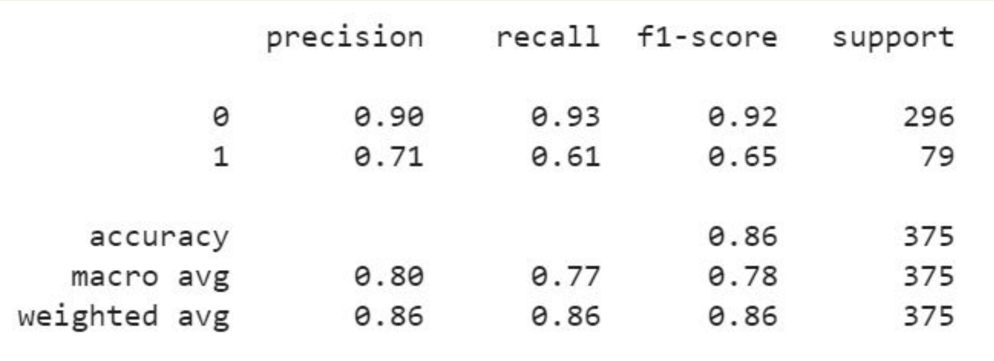
2. Support Vector Machine

In this project we used Support Vector Machine algorithm as a binary classifier. We divided our target into two groups:

I. Ratings above 7

II. Ratings below 7

We took a random sample of 1500 rows from our dataset to run this model on, since it was taking immense amount of time to run on the complete dataset.



```
              precision    recall  f1-score   support

           0       0.90      0.93      0.92       296
           1       0.71      0.61      0.65        79

    accuracy                           0.86       375
   macro avg       0.80      0.77      0.78       375
weighted avg       0.86      0.86      0.86       375
```

Benefits

1. SVM works relatively well when there is a clear margin of separation between classes.
2. SVM is more effective in high dimensional spaces.
3. SVM is effective in cases where the number of dimensions is greater than the number of samples.
4. SVM is relatively memory efficient

Limitation

1. SVM algorithm is not suitable for large data sets.
2. SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
3. In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
4. As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification

3. Random Forest Regression

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A random forest regressor works with data having a numeric or continuous output and they cannot be defined by classes.

Random Forest algorithm

- Pick at random k data points from the training set.

- Build a decision tree associated to these k data points.

- Choose the number N of trees you want to build and repeat steps 1 and 2.

- For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values.

Description Data splitting and Modelling

- Dependent and Independent variables were assigned to X(features except weighted average rating) and Y (weighted average rating) variables respectively

- Split the dataset into the Training set and Test set. The training set contains known output from which the model learns, test set then tests the model’s predictions based on what it learned from the training set, with the random state=78.

- Fitting the Standard Scaler with the training data. Scaling the data.

- Create a random forest regressor, did model fitting with n_estimator from 200 to 800 with the random state being 1

Results

Metrics for Regression Model used

1. Mean Squared Error
2. Root Mean Squared Error
3. Mean Absolute Error

Output for Estimators (200, 800, 100)

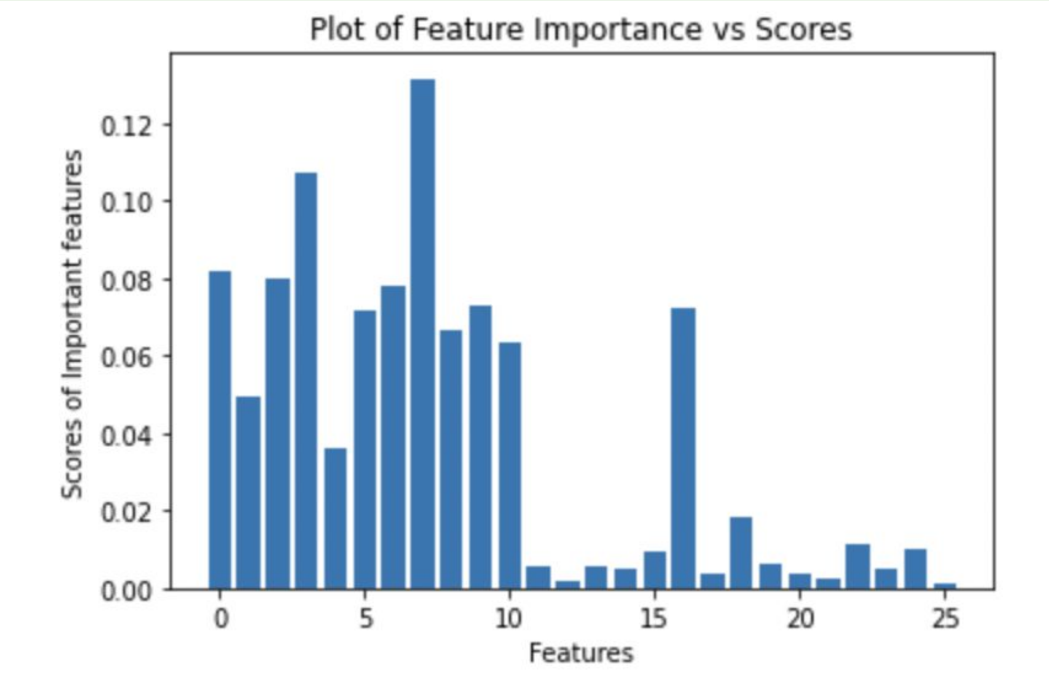
```
 [0.49415757514316994,0.4947008957127814, 0.4958920277984976,

0.4960607807872609,

0.4961116558077189,

0.4958916103852188]
```

The reason for the low score would be due to insufficent data relating to the average rating variable and relevance to the problem although the score has increased from previous linear regression model by 20% more.



Benefits

- Reduces overfitting in decision trees and improved performance

- Random Forests are not influenced by outliers to a fair degree. It does this by binning the variables

- Automates missing values present in the data

- Normalising of data is not required as it uses a rule-based approach.

Limitations

- Random forest is like a black box algorithm, you have very little control over what the model does

- Requires much time for training as it combines a lot of decision trees to determine the class

- Due to ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.

4. Gradient Boosting

Gradient boosting is a greedy algorithm and can overfit a training dataset quickly.

It can benefit from regularization methods that penalize various parts of the algorithm and generally improve the performance of the algorithm by reducing overfitting. Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set.

Benefits

- Often provides predictive accuracy that cannot be beat.
- Lots of flexibility - can optimize on different loss functions and provides several hyperparameter tuning options that make the function fit very flexible.
- No data pre-processing required - often works great with categorical and numerical values as is.
- Handles missing data - imputation not required.

Limitation

- GBMs will continue improving to minimize all errors. This can overemphasize outliers and cause overfitting. Must use cross-validation to neutralize.
- Computationally expensive - GBMs often require many trees (>1000) which can be time and memory exhaustive.
- The high flexibility results in many parameters that interact and influence heavily the behavior of the approach

5. Deep Learning

Deep learning is a subset of ML, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to “learn” from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy.

Benefits

- Features are automatically deduced and optimally tuned for desired outcome.
- Features are not required to be extracted ahead of time. This avoids time consuming machine learning techniques.
- Robustness to natural variations in the data is automatically learned.
- The same neural network based approach can be applied to many different applications and data types.

Limitation

- It requires very large amount of data in order to perform better than other techniques.
- It is extremely expensive to train due to complex data models.
- There is no standard theory to guide you in selecting right deep learning tools as it requires knowledge of topology, training method and other parameters

Summary of Results for Machine Learning Models

1. **Multiple linear Regression** Best Score : 0.244

2. **Support Vector Machine** Accuracy : 0.86

3. **Random Forest Regression** Best Score : 0.4961116558077189

4. **Gradient Boosting** Accuracy : 0.402757309246494394

5. **Deep Learning** Accuracy : **0.98**

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 Mean Score Error 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 AWS
9. Reports outcomes in Tableau for Visualization.
10. Final Summary

Thank You