Predicting IMDb Movie Ratings

Group 6: *Adegbenga Ayoola, Grace Le, Michelle Raj, Syed Bari, Wardah Anis*

ABOUT DATA

Data Source	Kaggle (IMDB Dataset)
Data Points	85855
Target	weighted_average_rating
Features	• Duration • Country • Language • Tot_voters_below_18 • Tot_voters_below_30 • Tot_voters_below_45 • Tot_voters_above_45 • Tot_male_voters • tot_female_voters

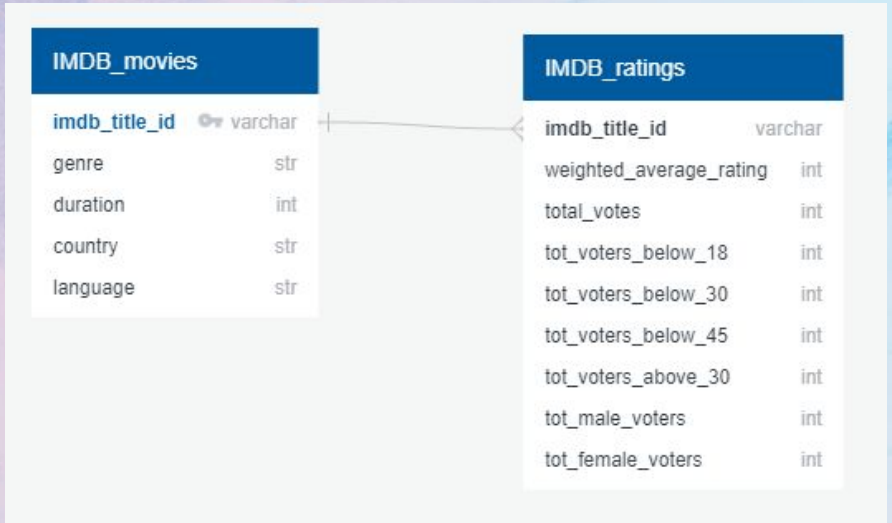
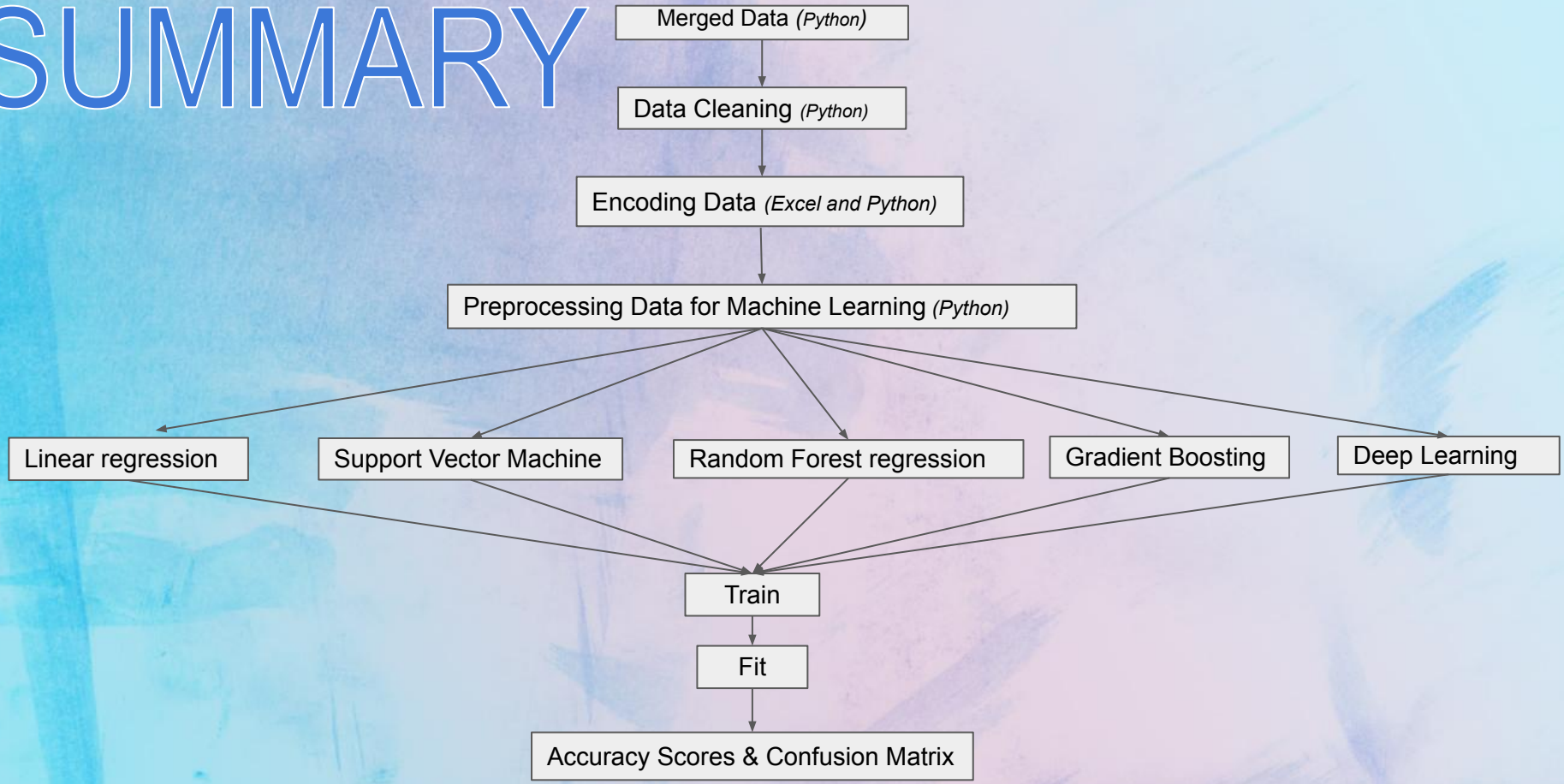


Figure 1: ERD Diagram

SUMMARY


Merged Data (Python)
Data Cleaning (Python)
Encoding Data (Excel and Python)
Preprocessing Data for Machine Learning (Python)
Linear regression
Support Vector Machine
Random Forest regression
Gradient Boosting
Deep Learning
Train
Fit
Accuracy Scores & Confusion Matrix

EXPLORATORY DATA ANALYSIS

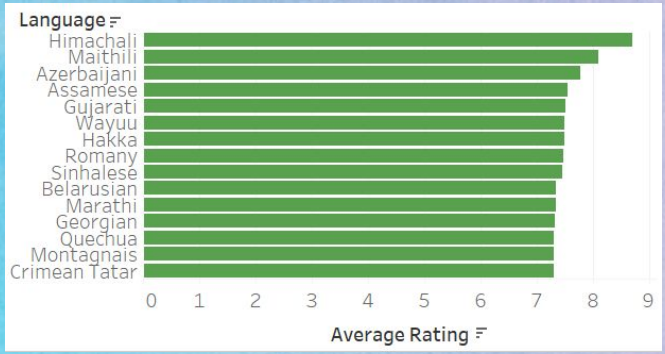


Figure 2: Top 15 languages with highest movie ratings

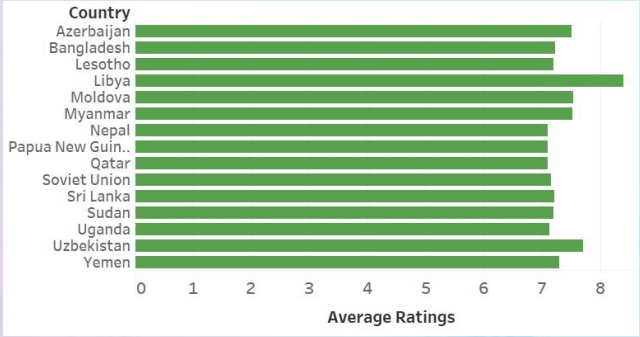


Figure 3: Top 15 countries that produces highest rated movies

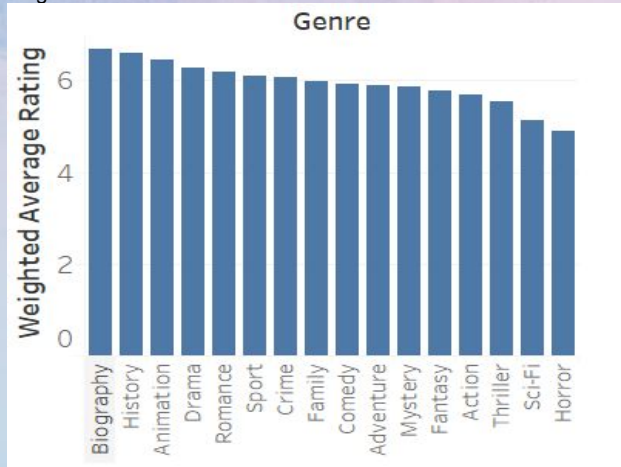
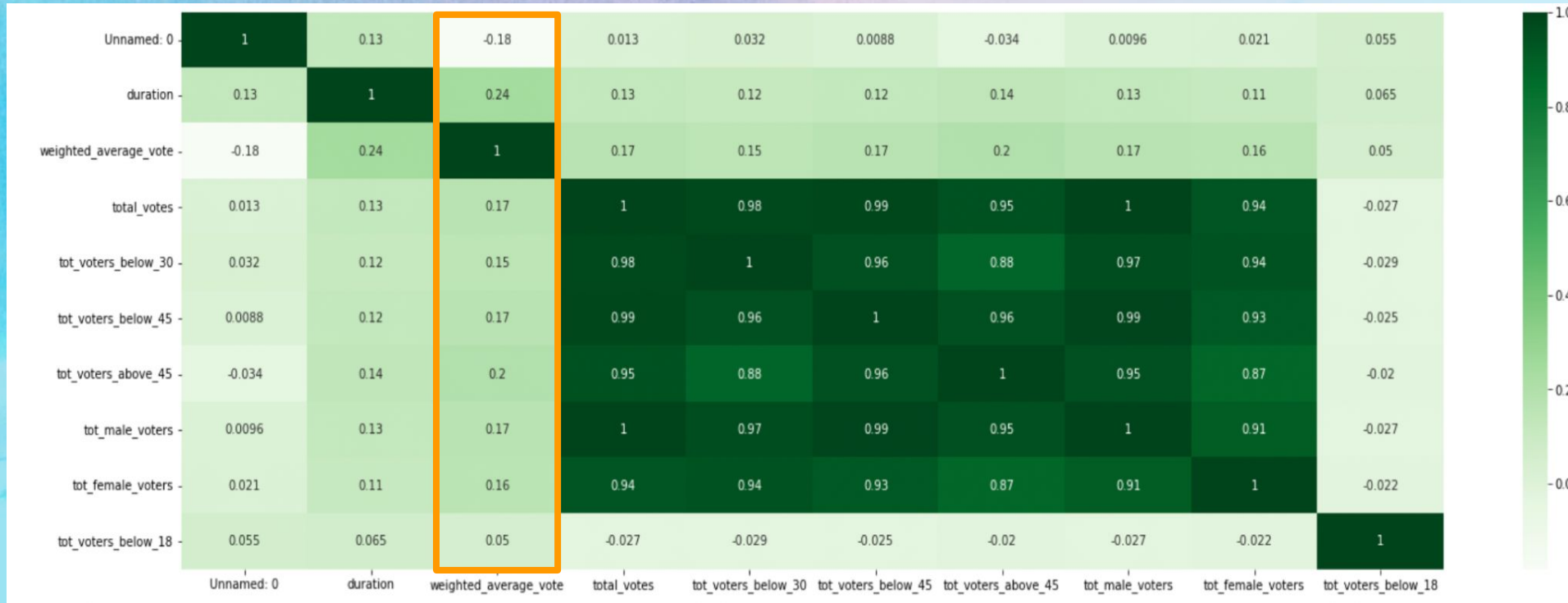


Figure 3: Genre vs Average Rating

Correlation Matrix for Cleaned Data



1. Multiple Linear Regression

```
r2 socre is  0.244894927271107 07
mean_sqrd_error is== 1.2087212257382316
root_mean_squared error of is== 1.0994185853159986
```

The reason for the low score would be due to insufficient 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.

2. Support Vector Machine

We divided our target into two groups:

I. Ratings above 7

II. Ratings below 7

	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

	Predicted Below 7	Predicted Above 7
Actual Below 7	276	20
Actual Above 7	31	48

3. Random Forest Regression

Description Data splitting and Modelling

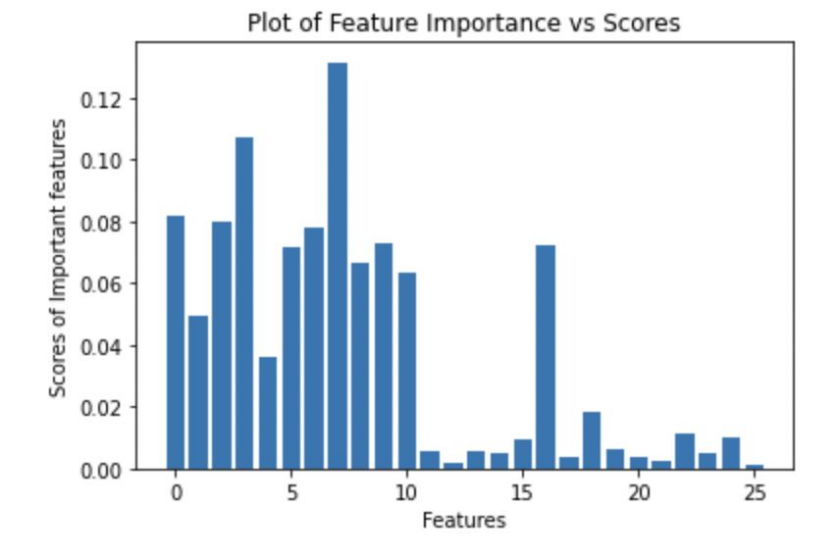
Metrics for Regression Model used

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

Output for Estimators (200, 800, 100)

[0.49415757514316994,0.4947008957127814,
0.4958920277984976,0.4960607807872609,0.4961116558077189,0.495891610385218

The reason for the low score would be due to insufficient 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.

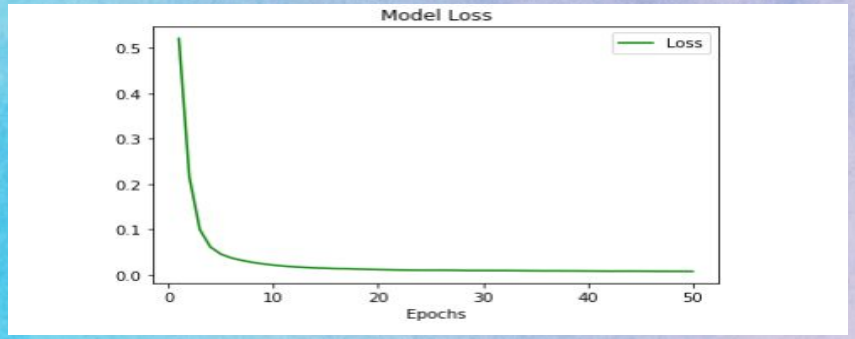


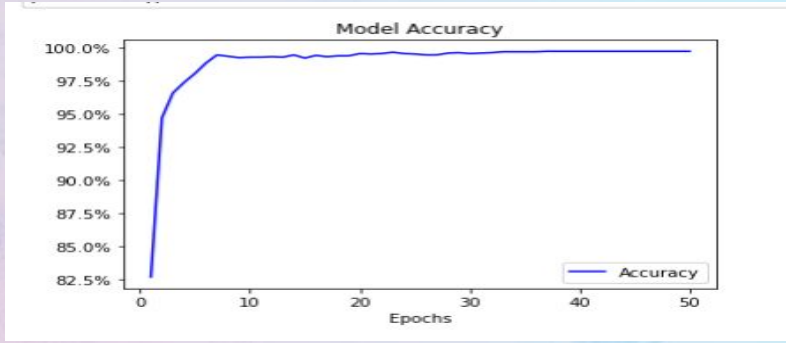
4. Gradient Boosting

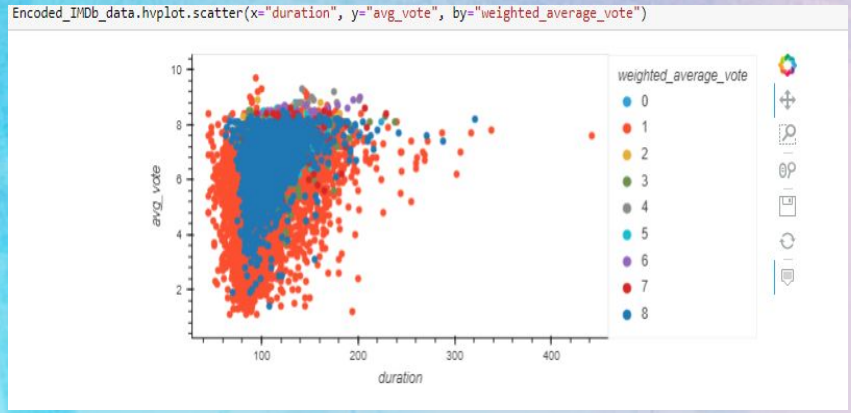
5 . Deep Learning

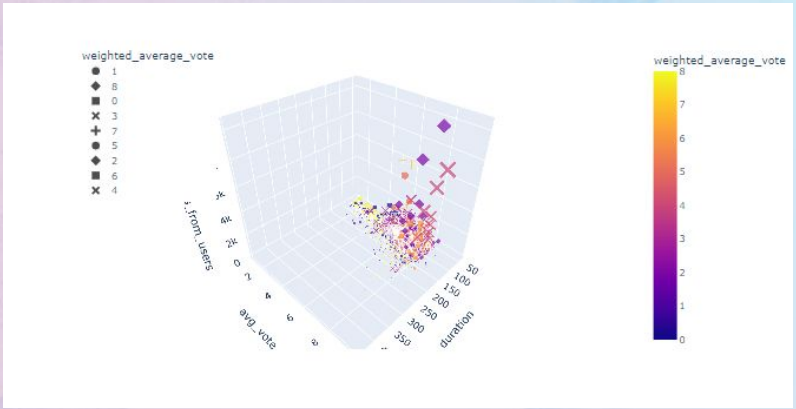
- They are an advanced form of machine learning that recognizes patterns and features in input data and provides a clear quantitative output.
- Used in classification algorithms and regression
- Detects complex non linear relationships
- Scalable and effective

5. Deep Learning cont.









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 : 0.50
4. **Gradient Boosting** Accuracy :0.40
5. **Deep Learning** Accuracy: 0.98

Recommendation for Future Analysis

1. Sentimental Analysis
2. Building Movie Recommendations System
3. Consider the Budget of the movie

What we would have done differently?

Utilized Dataset for Movie Ratings from various movie rating platforms and streaming services like Netflix, Rotten Tomatoes, Metacritic, Google to get more accurate prediction results

Thank You