VOTING PREDICTIVE MODEL

Anthony, Govarthini, Jayashree & Marina

PROJECT OVERVIEW

INTRODUCTION

IMDb (Internet Movie Database) is the world most popular online database of information related to films, television series, podcasts ratings, and fan and critical reviews.

INITIAL APPROACH

To implement machine learning models in order to predict accurately IMDB ratings of any particular movie.

After experimenting with the different regression models we found that none of them were giving the expected results for predicting IMDb rating score.

FINAL OBJECTIVE

We shift our attention into forecasting the number of votes for each film which is also an indicator of a film success in terms of popularity.

For this purpose, the relevant data that has been used is information about movies ratings, directors and star names, duration and other features that might influence the number of votes obtained by the public.

PROCESS OVERVIEW

1. Data Selection and Preparation

2. Round 1:
   a. Feature engineering
   b. Model testing

3. Round 2:
   a. Feature selection
   b. Model testing and final model selection

4. Final model optimization

1. Data Selection and Preparation

Dataset link: https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows/data

Original IMDb Rating data set : 16 columns x 1000 rows

- Data cleaning (casting data types, dropping null values, deleting columns)

Cleaned IMDb Rating data set : 14 columns x 715 rows

- Data exploration and visualization
- Pre Processed the data in order to implement ML models

Regroup categorical variables (mapping) - actors or film genre.

Scale numerical features

2a. Feature engineering and selection

- **Train test split**: Setting our target column (Number of votes) and pre-selecting the rest of the features that needed to be considered - Movie director, critic's score or film categories.

- **OneHotEncoder**: Utilized this function in order to categorize categorical values into numerical, enabling the performance of ML models.

- Normalize all the values by using the **MinMax Scaler**.

2b. Model Testing

The first round of model testing involved basic versions of the major regression machine learning models:

- KNN
- Linear Regression
- Decision Tree / Random Forest

In addition, basic ensemble approaches were utilised:

- Bagging and Pasting
- Ada boosting
- Gradient Boosting

The metrics used to compare models are:

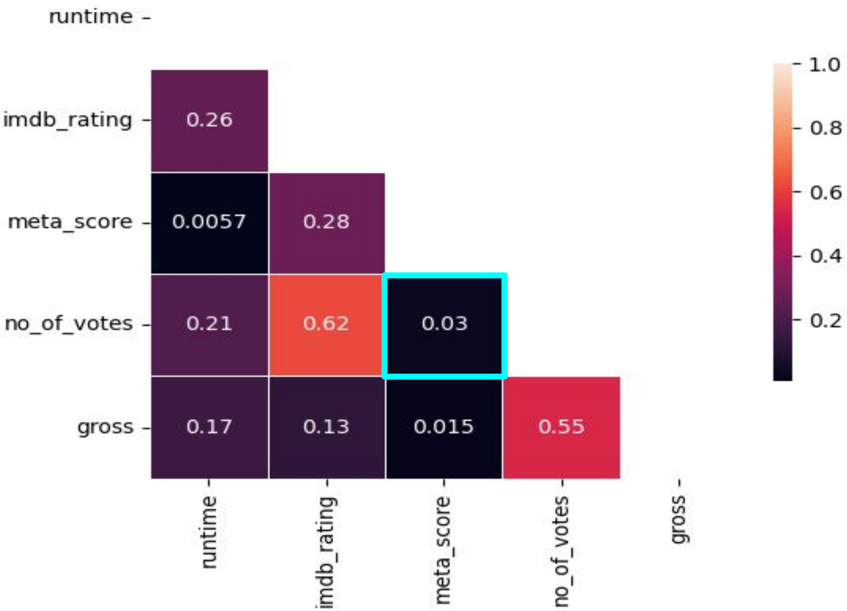
- r2_score
- mean_absolute_error
- mean_squared_error

3a. Feature Selection

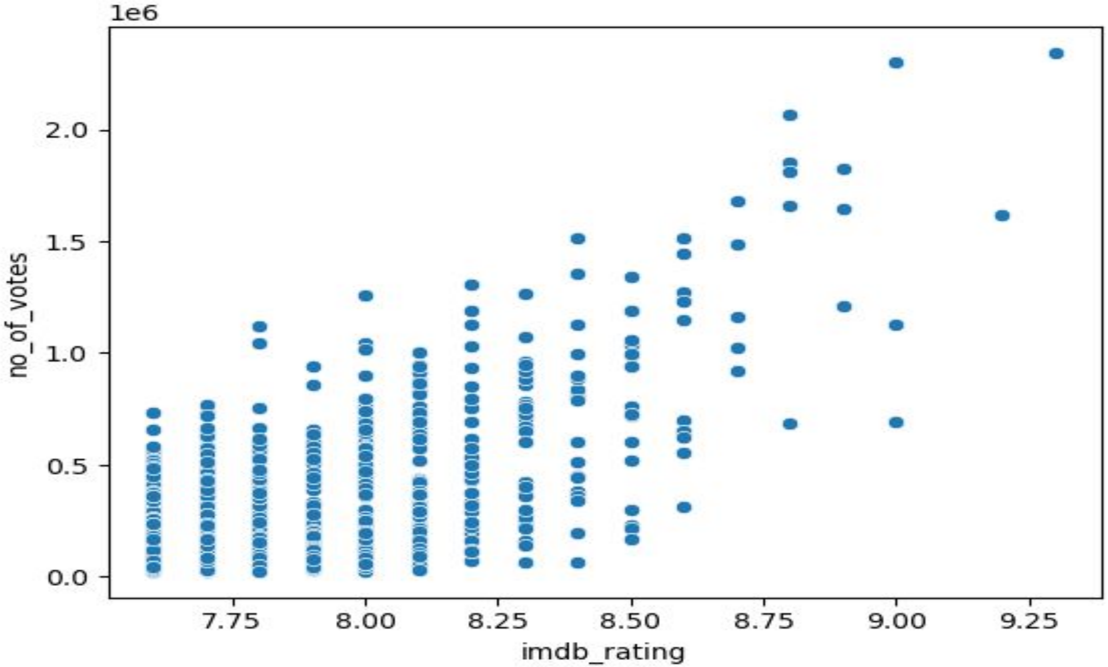
Initial approach was to utilize all the different features of the data set and perform trials with the different models to see the results.

After testing we dropped :

- **Meta_score** (critics) - low levels of correlation.
- **Star categories 2, 3 and 4** (casting) : Categorical values - discarded after testing.



A closer look at the Rating: Votes correlation



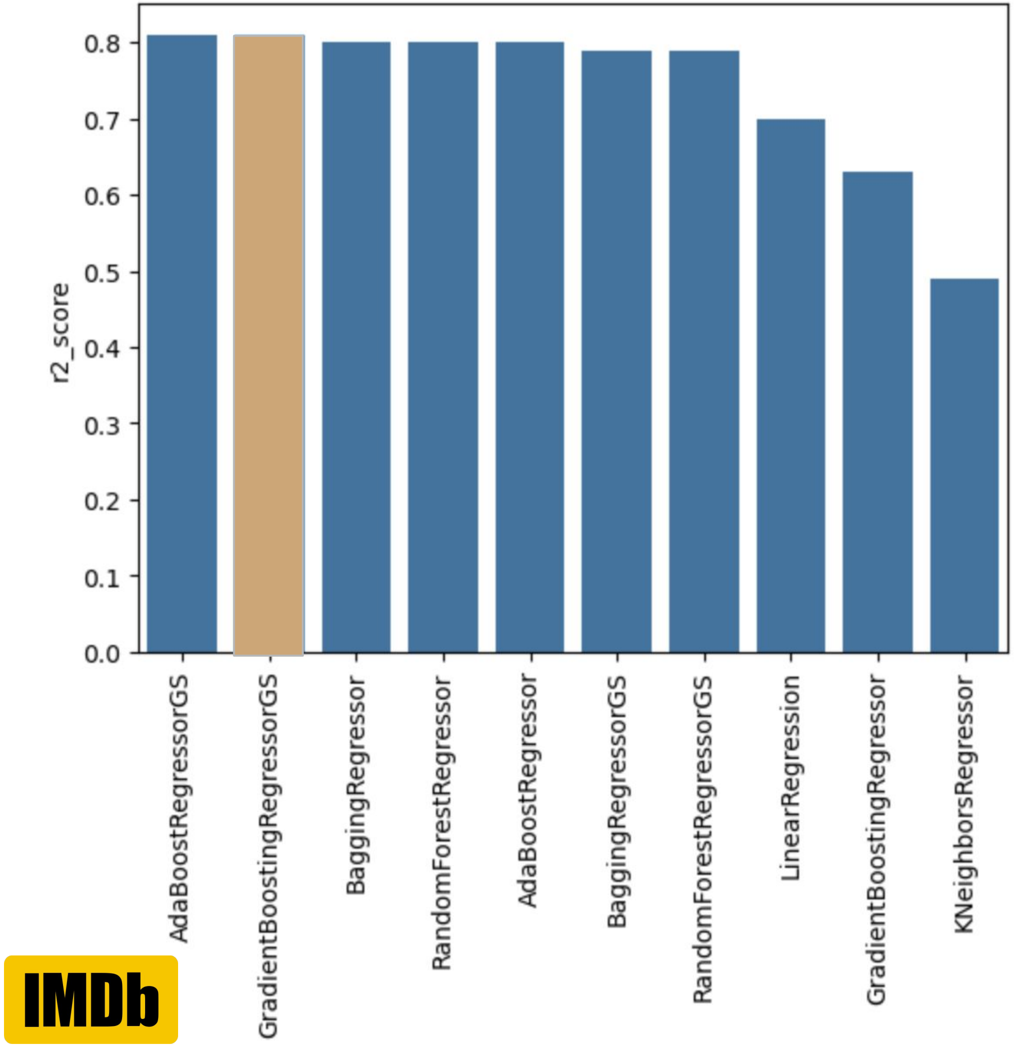
IMDb

3b. Model Testing and selection

- This involved the same steps and models used in the first round of model testing, but with a more refined set of features.
- In addition, initial parameter tuning was implemented.
- Ultimately, the GradientBoostingRegressor performed the best and was selected for final tuning.

3b. Model Comparison

- The GradientBoostingRegressor and AdaBoostRegressor performed equally on R2 Score, but Gradient Boosting performed better on the other metrics.



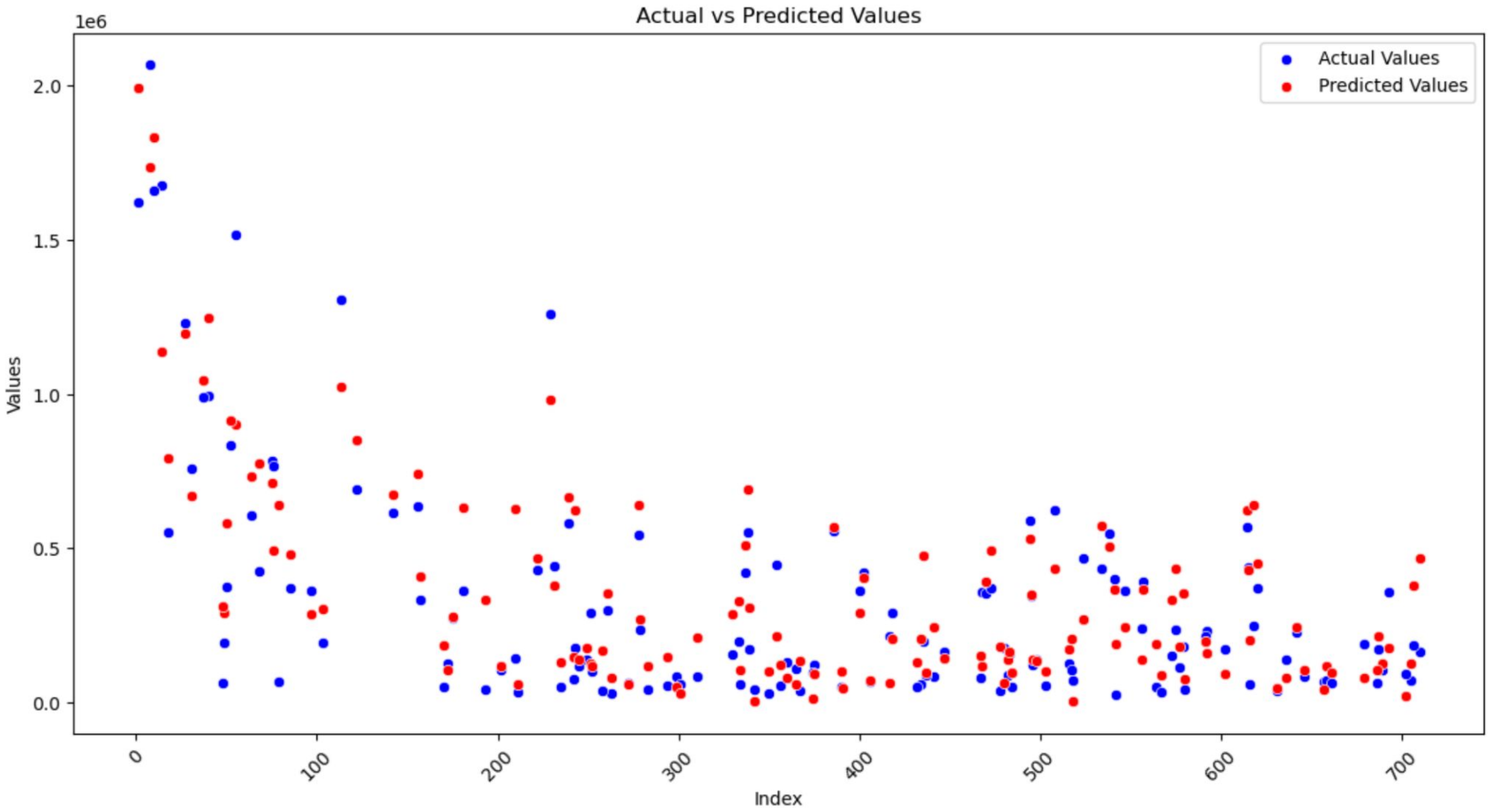
IMDb

4. Final model optimization

- Hyperparameter tuning technique: GridSearchCV

- Tuning process:
  - Run each adjusted model to retrieve metrics and best parameters
  - Add metrics and best parameters to a “scorecard” data frame for later comparison
  - Continually adjust parameters based on the results in the scorecard and repeat

- The final metrics:
  - R2 Score: **83%**
  - Mean Squared Error: 155,177.32
  - Mean Absolute Error: 110,789.33


Actual vs Predicted Values
1e6
2.0
1.5
1.0
0.5
0.0
Values
0
100
200
300
400
500
600
700
Index
● Actual Values
● Predicted Values

Challenges & Learnings

- **Data set:** Too many unique values in categorical columns
- **Initial approach:** Could not achieve highest rate for imdb_rating, ended up moving to no_of_votes column
- **Predictive models :** Understanding the relationships between the different features , how each model performs and interpreting the results.

Real-World Application and Impact

The global movies and entertainment market size was estimated at over USD 100 billion in 2023, expecting to grow at 8% during the current year.

Our model is developed on a real world data set and can also be used to predict user’s votes in other platforms such as Rotten Tomato or Metacritic. Other than films, TV shows or music shows popular votes could also be predicted using our model.

Forecasting the number of votes that a movie could obtain is a metric that may be also utilized to create a social media marketing strategy for a production company or streaming platform.

IMDb

IMDb

THANK YOU !

Anthony, Govarthini, Jayashree & Marina