OntarioTech
Engineering
& Applied Science

Movie Rating Prediction from IMDb Dataset

Presented by: Abeer Badawi and Mohamed Moutaoukil

Introduction

- In recent years, massive amounts of movies have become widely available for the general population, which hassled users and movie producers to deal with an overload of choices.
- Movie ratings in recent years have been influenced by various factors that make the accuracy prediction of ratings for new movies being released a complicated task.

Objectives and Goals

Predict the movie rating

- This analysis aims to identify the impact of the choices made during the production of a film. The goal is to predict a movie’s rating success based on a set of features before a movie is released.

Top movies common features

- We will use machine learning to discover what features have the most significant impact in determining rating and success.

Clients

- Help users to spend less time searching for a movie and more time watching.
- Help movie makers to decide if the movie will be successful or not.

Dataset

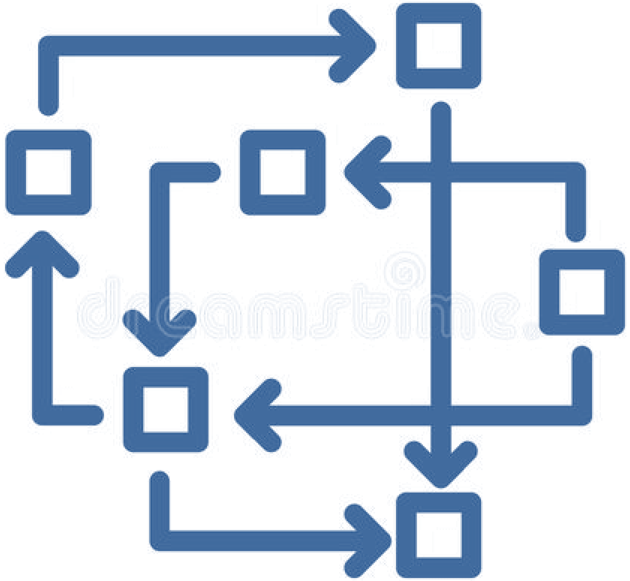
- The dataset includes information about several movies on IMDb, including movie titles, directors, genres, countries of origin, etc.
- Filtered the dataset to only keep movies that were produced in the year 2000 and after to remove bias of aged movies.
- Filtered the columns to only include the following features: Movie Title, IMDb Rating, Raters, Rated, Genres, Directors, Actors, Release Year, Budget, Revenue Worldwide, Production Company.

Methodology 1

- Narrowed down the movies that had at least 10,000 raters and a budget of at least 10,000 USD.
- Used one-hot encoder to convert the following category names of the dataset columns into values:
  - The “Rated” column
  - The “Genre” column
  - The “Directors” column but narrowed the selection down to only the top 10 directors
  - the “Actors” column but narrowed it down to the top 20 actors
  - The “Production Company” column but only kept the top 10 companies
- We added an extra column to the dataset and named it “Years Since Release” and created it by deducting the year the movie was released with the current year today.

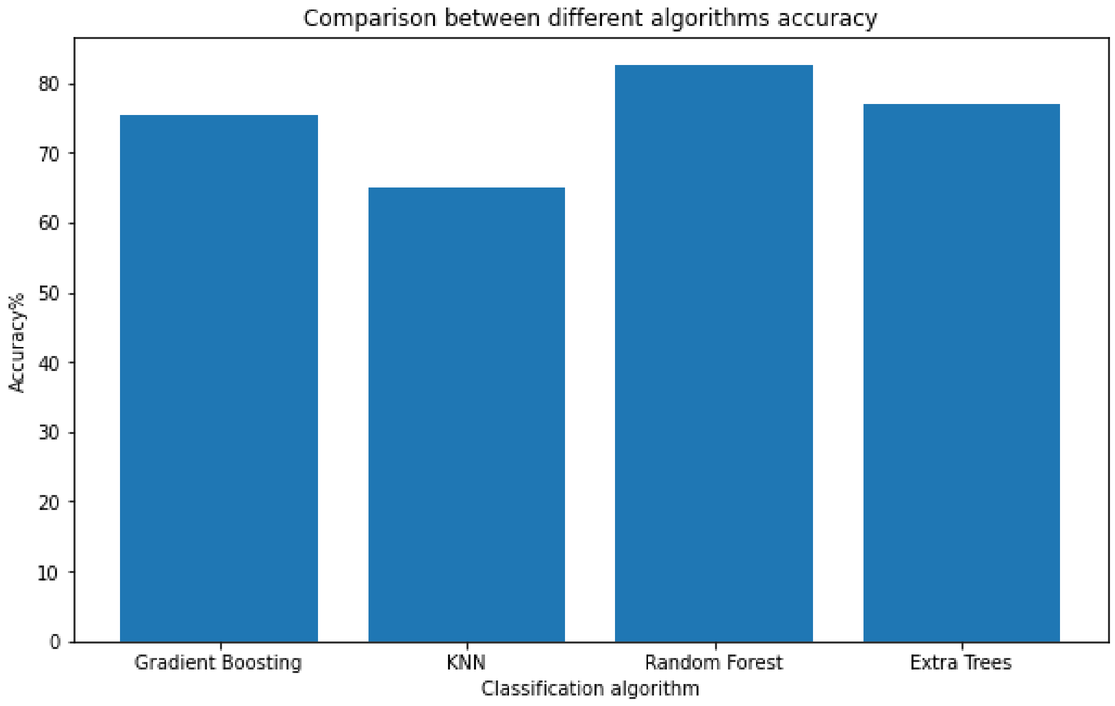
Methodology 2

- Classified the data by converting the IMDb ratings into three categories: category 0, category 1, category 2. The target labels were 0-4 for category 0, 4-7 for category 1, and 7-10 for category 2.
- We used 75% of the data as training data and the remaining 25% was used for training.
- utilized multiple data classifiers: GradientBoosting, Random Forest, KNN, and Extratrees.
- Chose the classifier that yielded the result with the highest accuracy

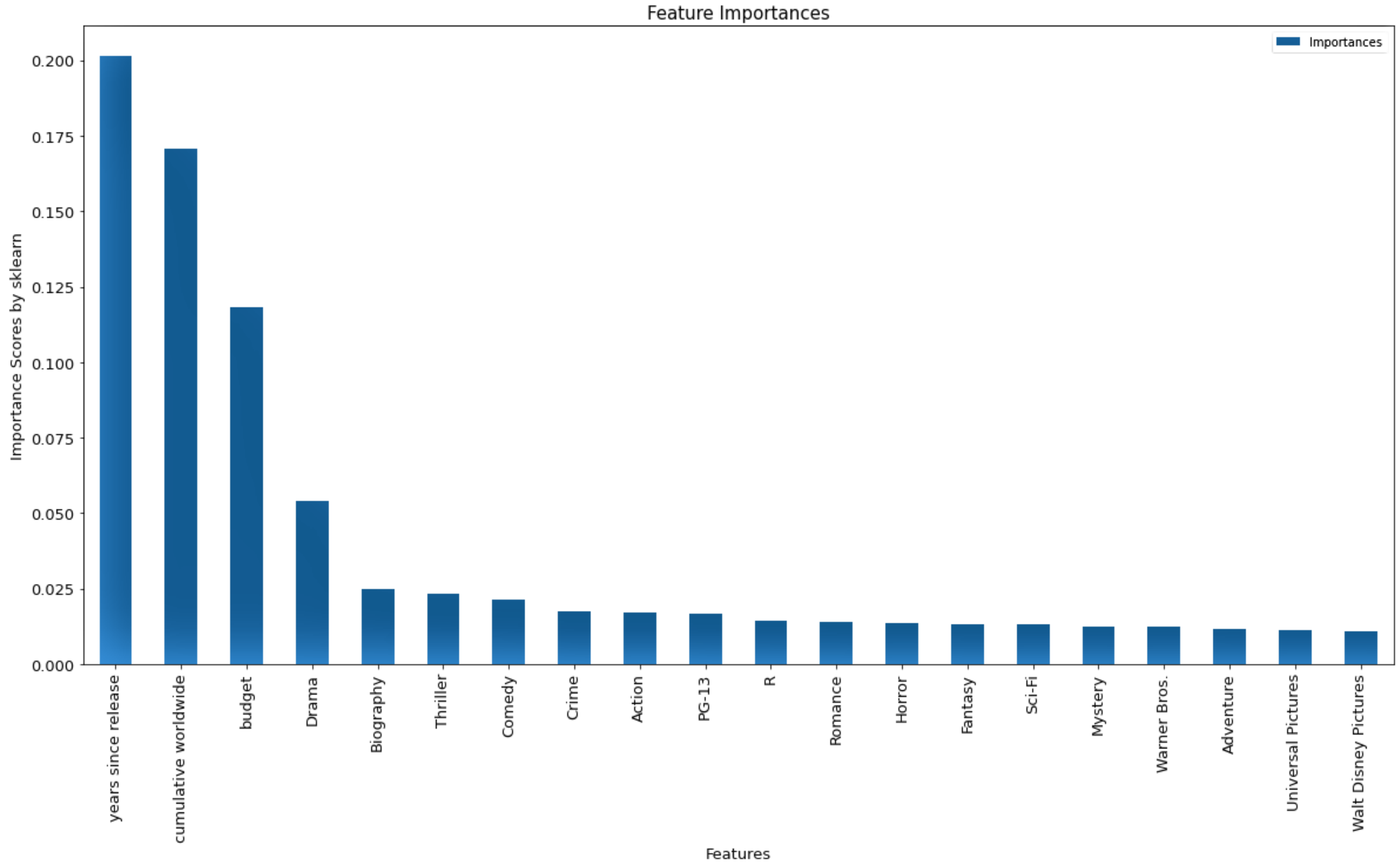


Results 1

- The four different classifiers we utilized obtained varying levels of accuracy:
  - ○ Gradient Boosting 75.30%
  - ○ Random forest 82.53%
  - ○ KNN 65.06%
  - ○ Extratrees 77.71%



Results 2



- We were able to describe the features that had the biggest impact on the success of a movie.
- The top 3 most important features are:
  1. Years since release
  2. Cumulative worldwide
  3. Budget
- The most popular genre is Drama

Results 3

- We can deduce the following conclusion about the predicted optimal features to create a successful movie (rating 7-10):

Genre	Director	Actor	Production Company	Rated
Drama	Christopher Nolan	Jake Gyllenhaal	Warner Bros.	PG-13
Biography	Guy Richie	Michael Fassbender	Universal Pictures	R
Thriller	Quentin Tarantino	Matthew McConaughey	Walt Disney Pictures	PG

Related Work

- S.Kabinsingha. proposed a new movie rating application based on machine learning algorithms with an accuracy of 80% using a decision tree algorithm on the IMDB dataset [10]. (Our algorithm has an accuracy of 82.53%)
- Zhou. found a movie rating predictive model with performance improvement with 5.91% from Netflix's Cinematch model.
- Rijul. provided an efficient approach to predict IMDb score on IMDb Movie Dataset. After building the five models, he found out that the Random Forest accurately predicted the movie rating at 61%. (Random Forest was our most accurate classifier as well at 82.53%)

Conclusion

- We were able to utilize Random forest classifier to create a movie rating prediction with 82.53% accuracy.
- Through the movie rating prediction engine, we were able to determine the optimal features for a successful movie.

Future Work

- Improve existing movie rating prediction classifiers to produce a higher accuracy percentage.
- Expanding selection of features defining our movie dataset for improved recommendation and predictive purposes

Thank you
Any questions?