**A collage of movie posters

Description automatically generated**

**MOVIE RECOMMENDER SYSTEMS**

**Final Report**

**December 2023**

# Table of Contents

[Table of Contents ii](#_Toc154102877)

[Table of Figures iv](#_Toc154102878)

[1. Overview 5](#_Toc154102879)

[1.1. The story of film 5](#_Toc154102880)

[1.2. Movie recommender systems 5](#_Toc154102881)

[2. The client 5](#_Toc154102882)

[3. The data 5](#_Toc154102883)

[4. Data collection 6](#_Toc154102884)

[5. Data wrangling 7](#_Toc154102885)

[5.1. Overview 7](#_Toc154102886)

[5.2. Conversion to csv files 7](#_Toc154102887)

[5.3. Removing unnecessary features 7](#_Toc154102888)

[5.4. Cleaning 7](#_Toc154102889)

[6. Exploratory data visualization and analysis 8](#_Toc154102890)

[6.1. Production countries 8](#_Toc154102891)

[6.2. Franchise movies 9](#_Toc154102892)

[6.3. Production companies 9](#_Toc154102893)

[6.4. Movie title wordcloud 10](#_Toc154102894)

[6.5. Original languages 10](#_Toc154102895)

[6.6. Popularity, vote average and vote count 11](#_Toc154102896)

[6.7. Movie release dates 13](#_Toc154102897)

[6.8. Spoken languages 17](#_Toc154102898)

[6.9. Runtime 17](#_Toc154102899)

[6.10. Budget 19](#_Toc154102900)

[6.11. Revenue 20](#_Toc154102901)

[6.12. Correlation matrix 22](#_Toc154102902)

[6.13. Genres 23](#_Toc154102903)

[6.14. Cast and crew 26](#_Toc154102904)

[7. Regression: Predicting movie revenues 29](#_Toc154102905)

[7.1. Feature engineering 30](#_Toc154102906)

[7.2. Model 30](#_Toc154102907)

[7.3. Feature importances 31](#_Toc154102908)

[8. Classification: Predicting movie success 31](#_Toc154102909)

[8.1. Model 31](#_Toc154102910)

[8.2. Feature importances 32](#_Toc154102911)

[9. Recommendation systems 32](#_Toc154102912)

[9.1. The simple recommender 32](#_Toc154102913)

[9.2. Content based recommender 33](#_Toc154102914)

[9.3. Collaborative filtering 35](#_Toc154102915)

[9.4. Hybrid recommender 36](#_Toc154102916)

[10. Conclusion 38](#_Toc154102917)

[References 39](#_Toc154102918)

# Table of Figures

[Figure 6.1. Production Countries for the MovieLens Movies (Apart from US). 8](#_Toc154102919)

[Figure 6.2. Movie title wordcloud 10](#_Toc154102920)

[Figure 6.3. Top 10 Original Languages for the MovieLens Movies (Apart from English). 11](#_Toc154102921)

[Figure 6.4. Distribution of Vote Average, Vote count - Vote average correlation 11](#_Toc154102922)

[Figure 6.5. Number of Movies released in a particular month 13](#_Toc154102923)

[Figure 6.6. Average Gross by the Month for Blockbuster Movies 14](#_Toc154102924)

[Figure 6.7. Month - Return boxplot 15](#_Toc154102925)

[Figure 6.8. Number of Movies released on a particular day 16](#_Toc154102926)

[Figure 6.9. Number of Movies each year 17](#_Toc154102927)

[Figure 6.10. Distribution of Runtime, Return - Runtime correlation 18](#_Toc154102928)

[Figure 6.11. Average of Runtime each year 18](#_Toc154102929)

[Figure 6.12. Distribution of Budget 19](#_Toc154102930)

[Figure 6.13. Budget - Revenue correlation 20](#_Toc154102931)

[Figure 6.14. Average of Revenue each year 21](#_Toc154102932)

[Figure 6.15. Correlation of numeric features of Movies 22](#_Toc154102933)

[Figure 6.16. Most popular genres of Movies 23](#_Toc154102934)

[Figure 6.17. Stacked Bar Chart of Movie Proportions by Genre. 24](#_Toc154102935)

[Figure 6.18. Line Chart of Movie Proportions by Genre. 24](#_Toc154102936)

[Figure 6.19. Genre – Revenue boxplot 25](#_Toc154102937)

[Figure 6.20. Casts with the Highest Total Revenue 26](#_Toc154102938)

[Figure 6.21. Directors with the Highest Total Revenue 27](#_Toc154102939)

[Figure 6.22. Casts with the Highest Average Revenue 28](#_Toc154102940)

[Figure 6.23. Directors with the Highest Average Revenue 29](#_Toc154102941)

[Figure 7.1. Regression model. 31](#_Toc154102942)

[Figure 8.1. Classification model. 32](#_Toc154102943)

[Figure 9.1. Top 250 recommended movies 33](#_Toc154102944)

## Overview

### The story of film

This section objective is narrating the history, trivia and facts behind the world of cinema through the lens of data. Extensive Exploration Data Analysis is performed on Movie Metadata about Movie Revenues, Casts, Crews, Budget, etc. through the years. Two predictive models are built to predict movie revenues and movie success. Through these models, what features have the most significant impact in determining revenue and success could be discovered later.

### Movie recommender systems

This part is concentrated on building multiple kinds of recommendation engines, named the Simple Generic Recommenders, the Content Based Filter and the User Based Collaborative Filter. The performance of the systems is evaluated in both a qualitative and quantitative manner.

## The client

The first section of the project does not have a definitive client. But some of the analysis performed in this part could be used in the Movie Making Business (Streaming Providers, Producers, etc). The Movie Succeed and Revenue Prediction Models can give valuable insights into the features that actually determine the end class and value respectively.

The Movie Recommender System is useful to any business that makes money via recommendations. This includes Amazon, Netflix, Hotstar, etc. Giving good recommendations directly entails one or many of the following:

1. Customers who buy a particular product or service leading to increased revenue or sales.
2. Customers who use the platform more frequently due to the quality and relevance of content shown to them.
3. Better user experience. Customers spend less time on searching and more time on watching. The pain of discovery is eliminated.

## The data

The data used in this project has been obtained from 2 sources: MovieLens and The Movie Database (TMDB).

MovieLens has a publicly available full dataset containing approximately 33,000,000 ratings and 2,000,000 tag applications applied to 86,000 movies by 330,975 users between January 09, 1995 and July 20, 2023. Includes tag genome data with 14 million relevance scores across 1,100 tags. This dataset was generated on July 20, 2023. A small subset of the dataset, containing 100,000 ratings and 3,600 tag applications applied to 9,000 movies by 600 users between March 29, 1996 and September 24, 2018. The subset was generated on September 26, 2018. [1]

One​ ​of​ ​the​ ​files​ ​contains​ ​the​ ​TMDB​ ​ID​ ​of​ ​every​ ​movie​ ​listed​ ​in​ ​the​ ​MovieLens​ ​dataset.​ ​Using​ ​this ID,​ ​the​ ​metadata,​ ​credits​ ​and​ ​keywords​ ​of​ ​all​ ​86,000​ ​movies​ ​were​ ​obtained​ ​by​ ​running​ ​a​ ​script that​ ​requested​ ​and​ ​parsed​ ​data​ ​from​ ​TMDB​ ​Open​ ​API.​ ​The​ ​data​ ​collected​ ​was​ ​initially​ ​in​ ​the JSON​ ​format​ ​but​ ​was​ ​converted​ ​into​ ​CSV​ ​files​ ​using​ ​Python’s​ ​Pandas​ ​Library. [2]

The​ ​following​ ​files​ ​were​ ​used​ ​in​ ​the​ ​project from both MovieLens subset database and TMDB:

1. **movies\_metadata.csv:**​ ​​The​ ​file​ ​containing​ ​metadata​ ​collected​ ​from​ ​TMDB​ ​for​ ​over 86,000​ ​movies.​ ​Data​ ​includes​ ​budget,​ ​revenue,​ ​date​ ​released,​ ​genres,​ ​etc.
2. **credits.csv:**​ ​​Complete​ ​information​ ​on​ ​credits​ ​for​ ​a​ ​particular​ ​movie.​ ​Data​ ​includes​ ​Director, Producer,​ ​Actors,​ ​Characters,​ ​etc.
3. **keywords.csv:**​ ​​Contains​ ​plot​ ​keywords​ ​associated​ ​with​ ​a​ ​movie.
4. **links\_small.csv:**​ ​​Contains​ ​the​ ​list​ ​of​ ​movies​.
5. **ratings\_small.csv:**​ ​​Contains ​100,000​ ​ratings​ ​on​ ​9,000​ ​movies from​ ​600​ ​users.​ ​The​ ​main​ ​dataset​ ​used​ ​for​ ​building​ ​the​ ​Collaborative​ ​Filter.

## Data collection

The​ ​MovieLens​ ​full and subset​ ​dataset​ ​is​ ​publicly accessible​ ​at​ ​the​ ​GroupLens​ ​[website](https://grouplens.org/datasets/movielens/latest/). The dataset includes in the file genome-scores.csv, genome-tags.csv, links.csv, movies.csv, ratings.csv and tags.csv. More details about the contents and use of all these files are in details in README.txt file.

TMDB [website](https://www.themoviedb.org/) suggests signing up for an API Key.​ ​This​ ​can allow individual​ ​access​ ​to​ ​data​ ​at​ ​3​ ​endpoints.​ ​Each​ ​endpoint gives​ ​details​ ​about​ ​the​ ​movie,​ ​its​ ​casts​ ​and​ ​crews​ ​information​ ​and​ ​plot​ ​keywords.​ ​Three separate​ ​scrapers were written​ ​to​ ​hit​ ​each​ ​endpoint​ ​and​ ​collect​ ​this​ ​data​ ​for​ ​all​ ​86,000​ ​movies.​ ​Since​ ​TMDB has​ ​a​ ​restriction​ ​in the 50 requests per second range,​ ​this​ ​task​ ​took​ ​a​ ​day​ ​to​ ​execute.

All​ ​the​ ​data​ ​collected​ ​was​ ​in​ ​the​ ​form​ ​of​ ​ ​JSON​ ​which​ ​demanded​ ​more​ ​processing

## Data wrangling

### Overview

This​ ​section​ ​describes​ ​the​ ​various​ ​data​ ​cleaning​ ​and​ ​data​ ​wrangling​ ​methods​ ​applied​ ​on​ ​the Movie​ ​datasets​ ​to​ ​make​ ​it​ ​more​ ​suitable​ ​for​ ​further​ ​analysis.​ ​The​ ​following​ ​sections​ ​are​ ​divided based​ ​on​ ​the​ ​procedures​ ​followed.

### Conversion to csv files

The​ ​data​ ​obtained​ ​from​ ​scraping​ ​was​ ​in​ ​the​ ​form​ ​of​ ​stringified​ ​JSON.​ ​This​ ​had​ ​to​ ​be​ ​converted into​ ​CSV​ ​Files​ ​to​ ​enable​ ​easier​ ​parsing​ ​and​ ​subsequent​ ​upload​ ​to​ ​public​ ​platforms​ ​such​ ​as Kaggle.

### Removing unnecessary features

Some​ ​features​ ​such​ ​as​ ​the​ ​Backdrop​ ​Path,​ ​Adult​ ​and​ ​IMDB​ ​ID​ ​were​ ​unnecessary​ ​attributes​ ​and were​ ​dropped​ ​to​ ​reduce​ ​the​ ​dimensions​ ​of​ ​the​ ​dataset.

### Cleaning

The​ ​dataset​ ​had​ ​a​ ​lot​ ​of​ ​features​ ​which​ ​had​ ​0s​ ​for​ ​values​ ​it​ ​did​ ​not​ ​possess.​ ​These​ ​values​ ​were converted​ ​to​ ​NaN.​ ​Some​ ​features​ ​were​ ​still​ ​in​ ​the​ ​form​ ​of​ ​a​ ​Stringified​ ​JSON​ ​Object.​ ​They​ ​were converted​ ​into​ ​Python​ ​Dictionaries​ ​using​ ​Python’s​ ​json​ ​library.​ ​These​ ​were​ ​further​ ​reduced​ ​into lists​ ​since​ ​we​ ​did​ ​not​ ​have​ ​a​ ​need​ ​for​ ​ID,​ ​timestamp​ ​and​ ​other​ ​attributes.

The​ ​dataframe​ ​was​ ​exploded​ ​wherever​ ​the​ ​analysis​ ​demanded​ ​it​ ​(for​ ​instance,​ ​genres​ ​and production​ ​countries).

Finally,​ ​most​ ​of​ ​the​ ​features​ ​were​ ​converted​ ​into​ ​a​ ​Python​ ​basic​ ​type​ ​(integer,​ ​string,​ ​float)​ ​byremoving​ ​all​ ​the​ ​unclean​ ​values.​ ​The​ ​date​ ​string​ ​was​ ​converted​ ​into​ ​a​ ​Pandas​ ​Datetime​ ​and​ ​fromit,​ ​the​ ​month,​ ​year​ ​and​ ​day​ ​of​ ​release​ ​of​ ​every​ ​movie was ​extracted​.

## Exploratory data visualization and analysis

In​ ​this​ ​section,​ ​the​ ​various​ ​insights​ ​produced​ ​through​ ​descriptive​ ​statistics​ ​and​ ​data​ ​visualizationis​ ​presented.

### Production countries

The Full MovieLens Dataset consists of movies that are overwhelmingly in the English language (more than 53,800). However, these movies may have shot in various locations around the world. It would be interesting to see which countries serve as the most popular destinations for shooting movies by filmmakers, especially those in the United States of America and the United Kingdom.

A map of the world

Description automatically generated

Figure 6.1. Production Countries for the MovieLens Movies (Apart from US).

Unsurprisingly, the United States is the most popular destination of production for movies given that the dataset largely consists of English movies.

Europe is also an extremely popular location with the UK, France, Germany and Italy in the top 5.

Japan and India are the most popular Asian countries when it comes to movie production.

### Franchise movies

The Star Wars Franchise is the most successful movie franchise raking in more than 8.785 billion dollars from 9 movies.

The James Bond Movies come in a separately second with a 7.816 billion dollars but the franchise has significantly more movies compared to the others which is 26 and therefore, a much smaller average gross.

In contrast, the The Avengers has only 4 movies but it grosses 7.776 billion dollars then it gets extremely high average gross.

### Production companies

Warner Bros is the highest earning production company of all time earning a staggering 78.3 billion dollars from close to 700 movies. Universal Pictures and 20th Century Fox are the second and the third highest earning companies with 77 billion dollars and 59 billion dollars in revenue respectively.

Marvel Studios has produced the most successful movies, on average. This is not surprising considering the amazing array of movies that it has produced in the last few decades: The Avengers, Iron man 3, Captain America: Winter Soldier, Avengers: Infinity War, Avengers: Endgame etc.

Pixar with an average gross of 551 million dollars comes in second with movies such as Up, Finding Nemo, Inside Out, Wall-E, Ratatouille, the Toy Story Franchise, Cars Franchise, etc. under its banner.

### Movie title wordcloud

A text in a rectangle

Description automatically generated

Figure 6.2. Movie title wordcloud

The word Love is the most commonly used word in movie titles. Girl, Man and Life are also among the most commonly occuring words. This encapsulates the idea of the ubiquitious presence of romance in movies pretty well.

### Original languages

There are 122 languages represented in the dataset. As expected, English language films form the overwhelmingly majority. French and Italian movies come at a very distant second and third respectively.

A graph of different colored rectangular shapes

Description automatically generated

Figure 6.3. Top 10 Original Languages for the MovieLens Movies (Apart from English).

As mentioned earlier, French and Italian are the most commonly occurring languages after English. Japanese and Hindi form the majority as far as Asian Languages are concerned.

### Popularity, vote average and vote count

A graph of a distribution of vote

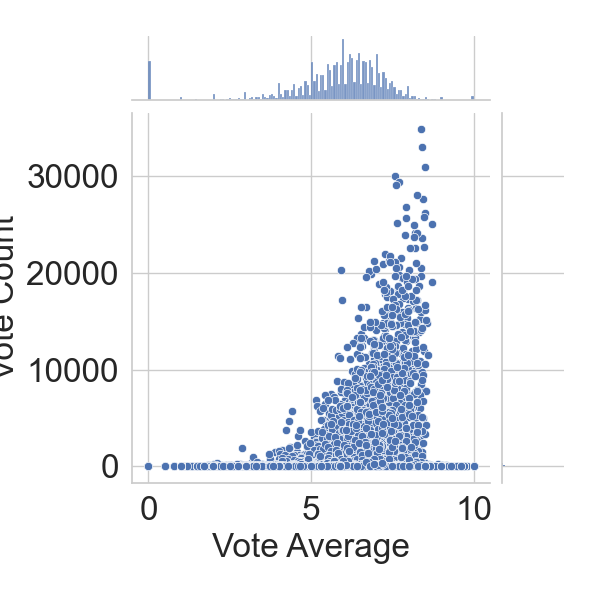
Description automatically generated

Figure 6.4. Distribution of Vote Average, Vote count - Vote average correlation

Oppenheimer is the most popular movie by the TMDB Popularity Score. Fast X and Mission: Impossible - Dead Reckoning Part One, two extremely successful action movies come in second and third respectively.

Inception and Interstellar, two critically acclaimed and commercially successful Christopher Nolan movies figure at the top of our chart.

It appears that TMDB Users are extremely strict in their ratings. The mean rating is only a 5.286 on a scale of 10. Half the movies have a rating of less than or equal to 6.

The Shawshank Redemption and The Dark Knight are the two most critically acclaimed movies in the TMDB Database. Interestingly, they are the top 2 movies in IMDB's Top 250 Movies list too. They have a rating of over 9 on IMDB as compared to their 8.7 and 8.5 TMDB Scores respectively.

There is a very small correlation between Vote Count and Vote Average. A large number of votes on a particular movie does not necessarily imply that the movie is good.

### Movie release dates

A graph of different colored bars

Description automatically generated

Figure 6.5. Number of Movies released in a particular month

It appears that October is the most popular month when it comes to movie releases.

Awards season: Many prestigious film awards, such as the Oscars, take place early in the following year. Studios release their best films towards the end of the year to be fresh in the minds of voters.

Holiday season: The holiday period, including Thanksgiving and Christmas, is a popular time for people to go to the movies. Studios release their most anticipated films during this time to take advantage of increased audience interest.

Box office performance: Historically, movies released in the fall and winter have performed well at the box office, as people tend to spend more time indoors during these months.

A graph of a number of movies

Description automatically generated with medium confidence

Figure 6.6. Average Gross by the Month for Blockbuster Movies

The months of April, May and June have the highest average gross among high grossing movies. This can be attributed to the fact that blockbuster movies are usually released in the summer when the kids are out of school and the parents are on vacation and therefore, the audience is more likely to spend their disposable income on entertainment.

A graph of different colored squares

Description automatically generated

Figure 6.7. Month - Return boxplot

The months of April, May and June have the highest average gross among high grossing movies. This can be attributed to the fact that blockbuster movies are usually released in the summer when the kids are out of school and the parents are on vacation and therefore, the audience is more likely to spend their disposable income on entertainment.

A graph of different colored bars

Description automatically generated

Figure 6.8. Number of Movies released on a particular day

Friday is clearly the most popular day for movie releases. This is understandable considering the fact that it usually denotes the beginning of the weekend. Sunday and Monday are the least popular days and this can be attributed to the same aforementioned reason.

A graph showing a line

Description automatically generated

Figure 6.9. Number of Movies each year

It is remarkably noticed that there is a sharp rise in the number of movies starting the 1990s decade. However, it is entirely possible that recent movies were oversampled for the purposes of this dataset.

### Spoken languages

The movie with the most number of languages, Train Station.

Train Station follows a single character, known only as "The Person in Brown", played by 40 actors who vary in age, gender, ethnicity and sexual orientation. Along the character's journey, they are presented with a series of choices - some minor, some life-altering. Cities include Berlin, Bogota, Dubai, Jakarta, Los Angeles, Singapore, Tehran and 20 others across five continents. Train Station unites cultures and breaks language barriers, reminding us that we all live in the same world full of diversity, options and consequences.

### Runtime

The average length of a movie is about 1 hour and 30 minutes. The longest movie on record in this dataset is a staggering 12,480 minutes (or 208 hours) long.

A graph of distribution of runtime

Description automatically generatedA graph of a running graph

Description automatically generated with medium confidence

Figure 6.10. Distribution of Runtime, Return - Runtime correlation

There seems to be relationship between the two quantities. The duration of a movie is independent of its success.

However, this might not be the case with duration and budget. A longer movie should entail a higher budget.

A graph showing the average of a stock market

Description automatically generated

Figure 6.11. Average of Runtime each year

It would be noticed that films started hitting the 60 minute mark as early as 1914. Starting 1924, films started having the traiditonal 90 minute duration and has remained more or less constant ever since.

It can be seen that every movie in this list were filmed in the late 1890s and the beginning of the 20th century. All these movies were one minute long.

Almost all the entries in the above chart are actually miniseries and hence, do not count as feature length films. It cannot gather too much insight from this list of longest movies as there is no way of distinguishing feature length films from TV Mini Series from the dataset (except, of course, by doing it manually).

### Budget

Budgets are expected to be a skewed quantity and also heavily influenced by inflation. Nevertheless, it would be interesting to gather as much insights as possible from this quantity as budget is often a critical feature in predicting movie revenue and success.

A graph with a number of bars

Description automatically generated with medium confidence

Figure 6.12. Distribution of Budget

The distribution of movie budgets shows an exponential decay. More than 75% of the movies have a budget smaller than 250 million dollars.

Avatar: The Way of Water film - the sequel to Avatar (2009) and the second installment in the Avatar film series - occupy the top spot in this list with a staggering budget of over 460 million dollars. All the top 10 most expensive films made a profit on their investment except for The Flash which managed to recoup about 90% of its investment, taking in a 264 million dollars on a 300 million dollar budget.

A graph with blue dots

Description automatically generated

Figure 6.13. Budget - Revenue correlation

Two quantities indicates a very strong correlation.

### Revenue

The mean gross of a movie is 52.8 million dollars whereas the median gross is much lower at 8.3 million dollars, suggesting the skewed nature of revenue. The lowest revenue generated by a movie is just 1 dollar whereas the highest grossing movie of all time has raked in an astonishing \*2.92 billion dollars.\*

A graph with blue lines and numbers

Description automatically generated

Figure 6.14. Average of Revenue each year

As can be seen from the figure, the maximum gross has steadily risen over the years. The world of movies broke the 1 billion dollar mark in 1997 with the release of Titanic. It took another few years to break the 2 billion dollar mark with Avatar. Both these movies were directed by James Cameron.

### Correlation matrix

A screen shot of a diagram

Description automatically generated

Figure 6.15. Correlation of numeric features of Movies

### Genres

A graph with different colored bars

Description automatically generated

Figure 6.16. Most popular genres of Movies

Drama is the most commonly occurring genre with almost half the movies identifying itself as a drama film. Comedy comes in at a distant second with more than 25% of the movies having adequate doses of humor. Other major genres represented in the top 10 are Thriller, Romance, Action, Documentary, Horror, Crime, Adventure, Family.

A bar chart with different colored bars

Description automatically generated

Figure 6.17. Stacked Bar Chart of Movie Proportions by Genre.

A line chart of movie proportions

Description automatically generated

Figure 6.18. Line Chart of Movie Proportions by Genre.

The proportion of movies of each genre has remained fairly constant since the beginning of this century except for Drama. The proportion of drama films has fallen by over 5%. Romance movies have enjoyed a astonishing increase in their share.

A graph of different colored squares

Description automatically generated

Figure 6.19. Genre – Revenue boxplot

Animation movies has the largest 25-75 range as well as the median revenue among all the genres plotted. Fantasy and Science Fiction have the second and third highest median revenue respectively.

### Cast and crew

A graph of various colored bars

Description automatically generated with medium confidence

Figure 6.20. Casts with the Highest Total Revenue

A graph of a company

Description automatically generated with medium confidence

Figure 6.21. Directors with the Highest Total Revenue

A graph of different colored bars

Description automatically generated

Figure 6.22. Casts with the Highest Average Revenue

A graph of a company

Description automatically generated with medium confidence

Figure 6.23. Directors with the Highest Average Revenue

## Regression: Predicting movie revenues

Predicting Movie Revenues is an extremely popular problem in Machine Learning which has created a huge amount of literature. Most of the models proposed in these papers use far more potent features than what possess at the moment. These include Facebook Page Likes, Information on Tweets about the Movie, YouTube Trailer Reaction (Views, Likes, Dislikes, etc.), Movie Rating (MPCAA, CBIFC) among many others.

To compensate for the lack of these features, there are going to be a few cheats. TMDB's Popularity Score and Vote Average will be being used as features in model to assign a nuerical value to popularity. However, it must be kept in mind that these metrics will not be available when predicting movie revenues in the real world, when the movie has not been released yet.

### Feature engineering

1. belongs\_to\_collection will be turned into a Boolean variable. 1 indicates a movie is a part of collection whereas 0 indicates it is not.
2. genres will be converted into number of genres.
3. homepage will be converted into a Boolean variable that will indicate if a movie has a homepage or not.
4. original\_language will be replaced by a feature called is\_english to denote if a particular film is in English or a Foreign Language.
5. production\_companies will be replaced with just the number of production companies collaborating to make the movie.
6. production\_countries will be replaced with the number of countries the film was shot in.
7. month will be converted into a variable that indicates if the month was a holiday season.
8. day will be converted into a binary feature to indicate if the film was released on a Friday.

### Model

The​ ​model​ ​chosen​ ​for​ ​regression​ ​is​ ​the​ ​Gradient​ ​Boosting​ ​Regression. Coefficient of Determination is 0.7213 which is a pretty score for the basic model that have built.

### Feature importances

A graph with a line

Description automatically generated with medium confidence

Figure 7.1. Regression model.

It has been ​noticed​ ​that​ ​vote\_count,​ ​a​ ​feature​ was ​cheated​ ​with,​ ​is​ ​the​ ​most​ ​important​ ​feature​ ​to​ the Gradient​ ​Boosting​ ​Model.​

​This​ ​goes​ ​on​ ​to​ ​show​ ​the​ ​importance​ ​of​ ​popularity​ ​metrics​ ​in determining​ ​the​ ​revenue​ ​of​ ​a​ ​movie.​ ​Budget​ ​was​ ​the​ ​second​ most​ ​important​ ​feature​ ​followed​ ​by Runtime ​and​ Popularity.

## Classification: Predicting movie success

As with the regression model, there are a few cheats and use features that may not be available in the real world for the lack of other useful popularity metrics.

Extensive analysis of data has already been performed and haven't been done a lot with respect to determining factors that make a movie a success. Attempt at doing that in this section and follow it up by building a model.

### Model

Gradient Boosting Classifier has an accuracy of 78%. Again, this model can be improved upon through hyperparameter tuning and more advanced feature engineering but since this is not the main objective of this project, this will be skipped.

### Feature importances

A graph with different colored bars

Description automatically generated

Figure 8.1. Classification model.

It is clear to see​ ​that​ ​vote​ \_count​ ​is​ ​once​ ​again​ ​the​ ​most​ ​significant​ ​feature​ ​identified​ ​by​ ​the ​ ​Classifier. Other​ ​important​ ​features​ ​include Year,​ ​Budget,​ Belongs to collection and Popularity.​ ​With​ ​this,​ ​it​ ​will​ ​conclude​ ​this discussion​ ​on​ ​the​ ​classification​ ​model​ ​and​ ​move​ ​on​ ​to​ ​the​ ​main​ ​part​ ​of​ ​the​ ​project.

## Recommendation systems

### The simple recommender

The Simple Recommender offers generalized recommendations to every user based on movie popularity and (sometimes) genre. The basic idea behind this recommender is that movies that are more popular and more critically acclaimed will have a higher probability of being liked by the average audience. This model does not give personalized recommendations based on the user.

The implementation of this model is extremely trivial. All have to do is sort movies based on ratings and popularity and display the top movies of the list. As an added step, pass in a genre argument to get the top movies of a particular genre.

The next step is to determine an appropriate value for \*m\*, the minimum votes required to be listed in the chart. Using 95th percentile as cutoff. In other words, for a movie to feature in the charts, it must have more votes than at least 95% of the movies in the list.

An overall Top 250 Chart is built and defined a function to build charts for a particular genre.

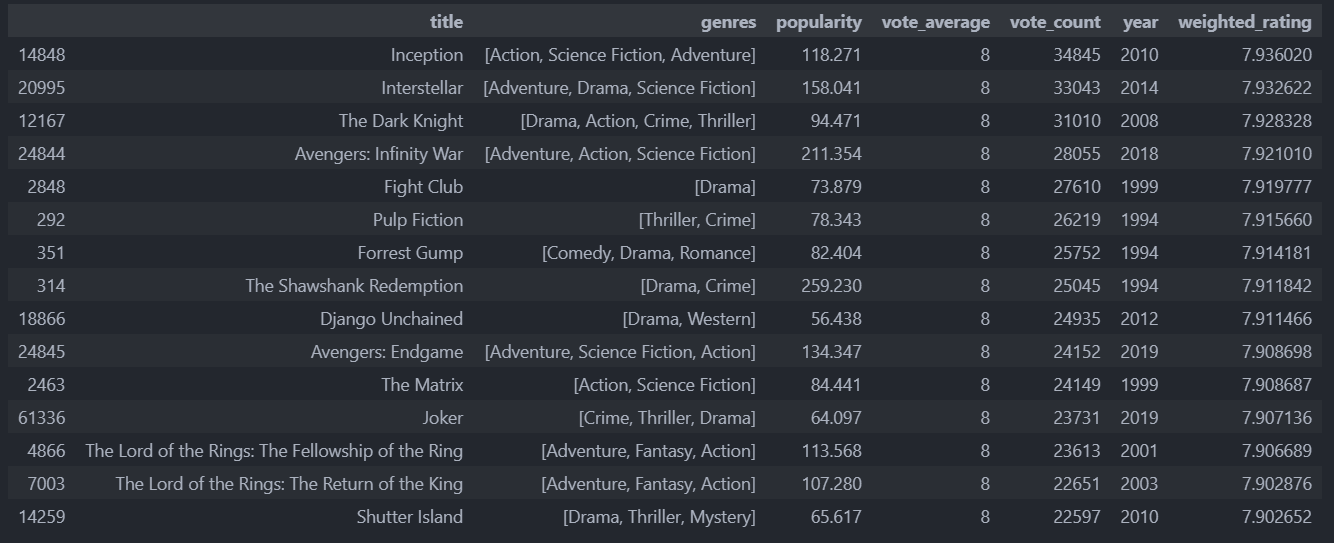


Figure 9.1. Top 250 recommended movies

Three Christopher Nolan Films, Inception, Interstellar and The Dark Knight occur at the very top of the chart. The chart also indicates a strong bias of TMDB Users towards particular genres and directors.

### Content based recommender

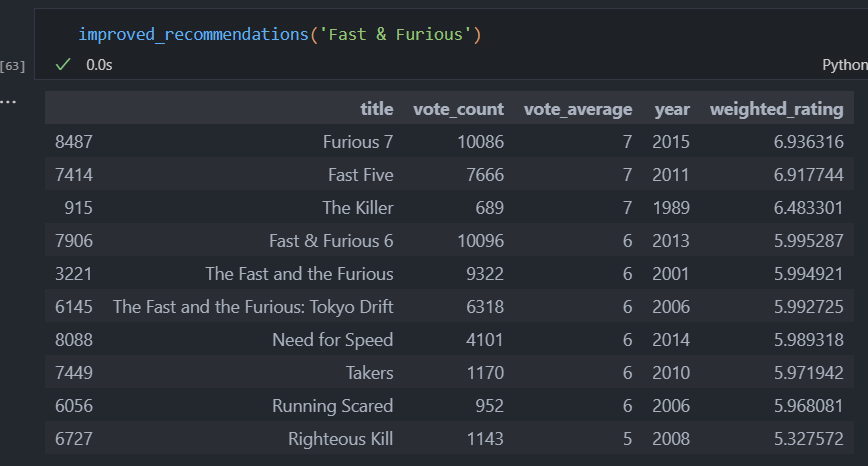
The recommender built in the previous section suffers some severe limitations. For one, it gives the same recommendation to everyone, regardless of the user's personal taste. If a person who loves romantic movies (and hates action) were to look at our Top 15 Chart, she/he wouldn't probably like most of the movies. If she/he were to go one step further and look at the charts by genre, she/he wouldn't still be getting the best recommendations.

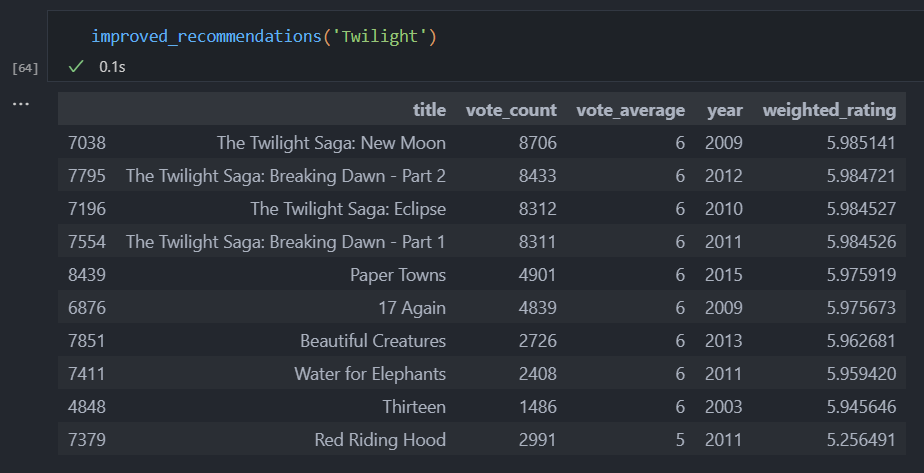
To personalise recommendations more, an engine is going to be built that computes similarity between movies based on certain metrics and suggests movies that are most similar to a particular movie that a user liked. Since using movie metadata (or content) to build this engine, this also known as Content Based Filtering.

Build two Content Based Recommenders based on:

* Movie Overviews and Taglines
* Movie Cast, Crew, Keywords and Genre

Also, as mentioned in the introduction, using a subset of all the movies available due to limiting personal computing power available.





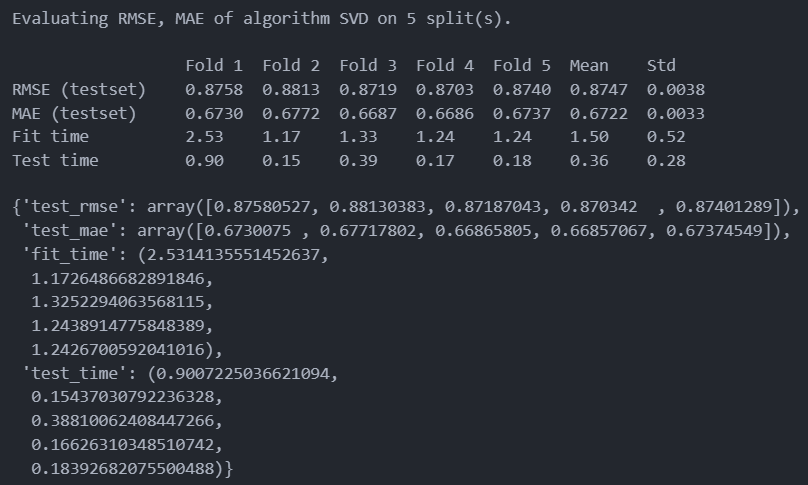
### Collaborative filtering

The Content Based engine suffers from some severe limitations. It is only capable of suggesting movies which are \*close\* to a certain movie. That is, it is not capable of capturing tastes and providing recommendations across genres.

Also, the engine that built is not really personal in that it doesn't capture the personal tastes and biases of a user. Anyone querying this engine for recommendations based on a movie will receive the same recommendations for that movie, regardless of who she/he is.

Therefore, in this section, it is about to use a technique called Collaborative Filtering to make recommendations to Movie Watchers. Collaborative filtering filters information by using the interactions and data collected by the system from other users. It’s based on the idea that people who agreed in their evaluation of certain items are likely to agree again in the future.

It will not be implementing Collaborative Filtering from scratch. Instead, it will use the Surprise library that used extremely powerful algorithms like Singular Value Decomposition (SVD) to minimise RMSE (Root Mean Square Error) and give great recommendations.

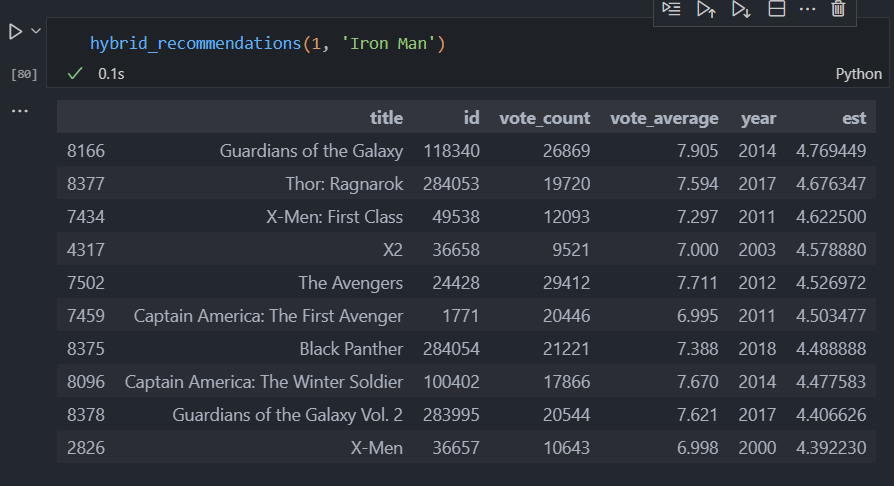


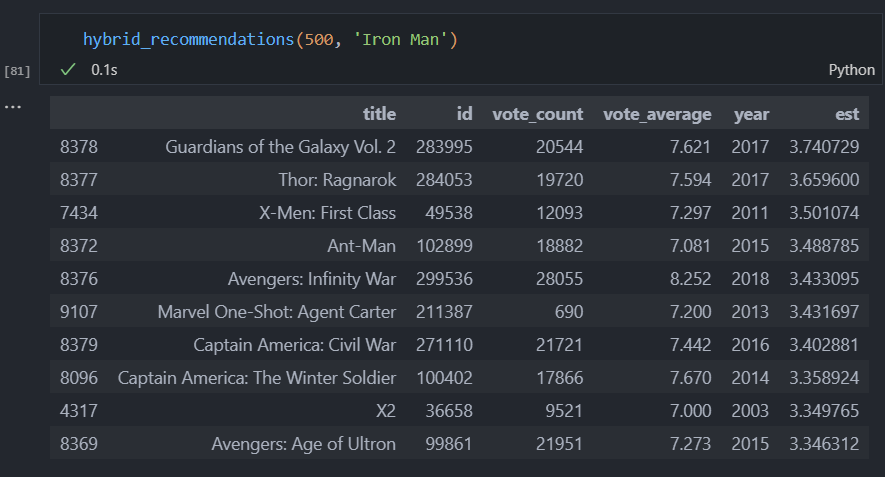
A mean Root Mean Sqaure Error of 0.8747 which is more than good enough for this case.

### Hybrid recommender

In this section, a main purpose is to build a Simple Hybrid Recommender that brings together techniques that have been implemented in the Content Based and Collaborative Filter Based engines. This is how it will work:

* Input: User ID and the Title of a Movie
* Output: Similar movies sorted on the basis of expected ratings by that particular user.





It can be seen that for the Hybrid Recommender, it can get different recommendations for different users although the movie is the same. Hence, this recommendations are more personalized and tailored towards particular users.

## Conclusion

In this notebook, 4 different recommendation engines have been built based on different ideas and algorithms. They are as follows:

1. **Simple Recommender:** This system used overall TMDB Vote Count and Vote Averages to build Top Movies Charts, in general and for a specific genre. The IMDB Weighted Rating System was used to calculate ratings on which the sorting was finally performed.
2. **Content Based Recommender:** This system was built by two content based engines: One that took movie overview and taglines as input. The other which took metadata such as cast, crew, genre and keywords to come up with predictions. They were also deviced a simple filter to give greater preference to movies with more votes and higher ratings.
3. **Collaborative Filtering:** The powerful Surprise Library was used to build a Collaborative Filter based on single value decomposition. The RMSE obtained was less than 1 and the engine gave estimated ratings for a given user and movie.
4. **Hybrid Engine:** All the ideas from Content and Collaborative Filterting were brought together to build an engine that gave movie suggestions to a particular user based on the estimated ratings that it had internally calculated for that user.

The​ ​code​ ​associated​ ​with​ ​this​ ​report​ ​is​ ​available​ ​at:

https://github.com/thanhnghth99/recommender\_system.git

# References

|  |  |
| --- | --- |
| [1] | Department of Computer Science and Engineering at the University of Minnesota, "MovieLens," 1997. [Online]. Available: https://grouplens.org/datasets/movielens/. [Accessed 01 December 2023]. |
| [2] | "The Movie Database (TMDB)," 2008. [Online]. Available: https://www.themoviedb.org/. [Accessed 01 December 2023]. |