Movie Recommender System Using Collaborative Filtering Technique in PySpark

BY

STUDENT ID: P2679181

Name: Abiodun Samuel Ogundairo

Table of Contents

Abstract ..................................................................................................................................................3

1.1 Introduction ................................................................................................................................. 3

2.1 Brief Literature ............................................................................................................................. 1

2.1.1 Research Questions Formulation ............................................................................................ 1

2.1.2 Justification of the Study ......................................................................................................... 1

3.1 Dataset Description ..................................................................................................................... 1

3.2 Data exploration .......................................................................................................................... 2

3.3 Data modelling ............................................................................................................................. 2

3.3.1 Environment Set up ................................................................................................................. 2

3.3.2 Modelling ................................................................................................................................. 3

3.3.3 Evaluating Selected Model on a user ...................................................................................... 4

3.3.4 Movie rating simulation ........................................................................................................... 4

4.1 Conclusion .................................................................................................................................... 5

4.1.1 Challenges and lessons learned ............................................................................................... 5

**References** ............................................................................................................................................. 6

**Abstract**

The application of model-based collaborative filtering technique, which helps to personalized movies to users based on their profile and similarity with other users, is crucial because users find it challenging to scroll through thousands of recently released movies. Movies have served as a means of entertainment and have also contributed to the global economy. Therefore, the purpose of this project is to investigate movies through the lens of 20 million datasets and estimate user ratings on unwatched movies using big data analytics technologies such as PySpark.

# 1.1 Introduction

A recommendation system is a kind of information filtering that shows consumers information based on their profile information or previously expressed interests for particular products. As a result of the exponential growth in the number of movies made each year, the human brain is unable to comprehend or memorize it all at once without becoming confused. As a result, there is a need for automation and selection that may assist process information and tailor movies for viewers (Ponnam L.T et al., 2016). Companies that have used the recommendation system include Facebook, which uses it to suggest friends based on their networks of friends, Amazon, which suggests books and other products based on previous purchases, Google, which bases its recommendations on previous searches, Youtube, which bases them on previously watched videos, and Movie Lens, which in this study recommends movies to users based on preferences using collaborative filtering of viewer ratings and reviews on particular films. The matching of an item based on a user's profile and item information is known as content-based matching. Collaborative filtering, which extracts item information or patterns based on user cooperation or item similarity across users, is known as item information or pattern extraction. This work used the model-based method of collaborative filtering to construct a recommendation system for movie rating using the Movie Lens dataset. The collaborative system consists of two techniques: the neighbourhood and model-based technique (Deshpande M. et al., 2004). A study by Wen Z. (2008) compared the performance of the K nearest neighbor (KNN), Naive Bayes, and Sparse singular value decomposition (SVD) algorithms using user-submitted items from the Netflix dataset. The study found that Sparse SVD performed better in comparison because it had the lowest root mean squared error (RMSE), and it recommended that future work apply different algorithms to the dataset in the hopes of achieving a much lower RMSE. Everyone may now leave a digital imprint on the internet because of how the world is now connected through social media platforms, it allows users to express their thoughts, sentiments, and emotions. These platforms acquired a ton of data that can be examined, analyzed, and predicted in a variety of circumstances, from the most straightforward to the most complicated. One of the social media sites that offer movie information as well as space for user reviews, ratings, and votes is Movie Lens. The movie industry supports marketing and worldwide revenue in addition to entertainment and leisure. The discussion and rating of movies on social media platforms, however, gives data analysts the opportunity to gather information for exploration, analysis, and prediction of movie success because people's perspectives on movies vary depending on the region, necessitating the need for data that captures all pertinent information on a platform (such as IMDb, Metacritic, or Rotten Tomatoes). However, movie success is as important to directors and other officials as the production itself.

# 2.1 Brief Literature

In a research by Gaenssle, S. et al. (2018), factors including actor, script, actress, and directors were used to predict the success of foreign films in Russia. Saraee et al. (2004) also employed the same features in their data mining method for movie analysis and prediction. According to Jack Valenti, President and CEO of the Motion Picture Association of America (MPAA), no one can accurately predict the factors that affect a movie's commercial success; as a result, the film industry has tapped into the power of data in areas like movie success prediction with information available before the release and recommendation models. In the past, three aspects were used to determine if a movie was a success or a failure: brand-related (the franchise and the stars), evaluation-related (the ratings and the critics), and distribution-related (the budget and the copies) (Gaenssle S et al., 2018) This are all evaluated after a film has been made, but with the use of statistical models, movie ratings may be used to forecast a film's success. The success of a movie may depend on a number of factors; for example, some films may be considered successful based on the amount of money they make, while others may be determined by the reviews and ratings they receive (Quader, N et al., 2017).

## 2.1.1 Research Questions Formulation

1. What techniques of data exploration techniques were used?
2. Which performance metrics technique was used to evaluate the model?
3. Which parameter tuning improves model performance?

## 2.1.2 Justification of the Study

# This study is important since it investigates the Alternating Least Square machine learning framework in Spark in the creation of a recommendation system with the capacity to predict movie ratings. Despite the fact that this study will add to the body of literature that has already used machine learning techniques to design recommender systems for movie ratings, it will also serve as a framework or platform for movie directors and other officials to prototype the movie production before it is released to the public, perhaps giving them a preview of what to expect from their work.

# 3.1 Dataset Description

The 20 million records from the dataset utilized for this study were chosen and extracted into the following spreadsheet in comma-separated values (CSV) format from [Movie Lens](https://grouplens.org/datasets/movielens/latest/).

Movies: There are 58098 rows of movieId, title, and genre columns.

Ratings: It has 27753444 rows of movieId, userId, rating, and timestamp fields.

UserId, tags, movieId, and timestamp columns make up its 1108997 rows under the heading "Tags."

Links: It has 58098 rows of movieId, ImdbId, and tmdbId columns.

Genome-scores: It has 1486358 rows and the fields movieId, tagId, and significance.

Genome-tag: It has 1128 rows of tag Id and tag columns.

# As will be disclosed in the modelling section, the links and movies spreadsheets were utilized to condense prediction exploration while Tags, Genome-scores, and Genome-tags were ignored. The rating spreadsheet was used for training and verifying the collaborative filtering predictive model.

# 3.2 Data exploration

The dataset did not contain any missing values or outliers, hence the major preprocessing step used in this study was to remove the unnecessary timestamp column as shown in figure 1.

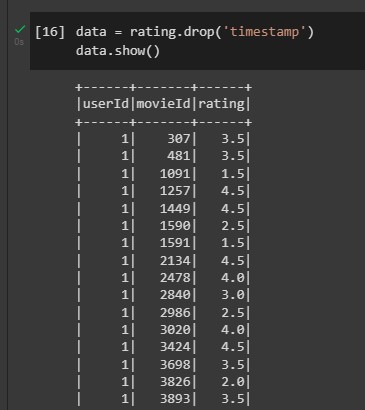


Figure 1: Data exploration

The pixie dust function was deployed to visually explore the dataset but was unable to provide visuals as such the matplotlib function was used to visually explore the dataset. The datasets were presented in a data frame format prior to the exploration as shown below.

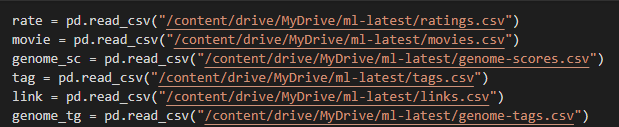


Figure 2: Data frame of datasets

The figures below showed exploration of the datasets.

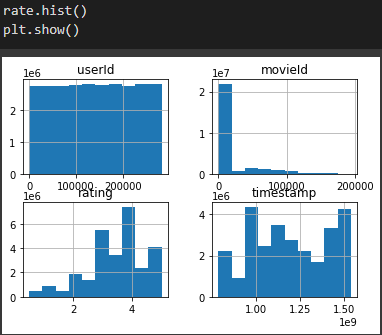
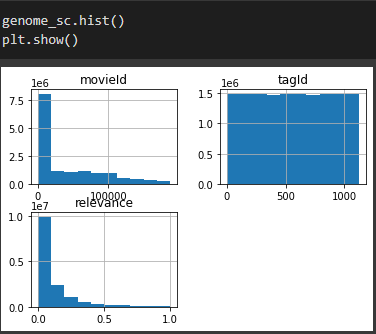


Figure 3: Histogram of rating

The visual showed that most movies range mostly from 2.5 and 4.0.



The relevance score showed the importance of a tag to a movie, with a 1 indicating strongly relevant and 0 indicating less relevant tag. As shown in the figure above the relevance score of movies tag in the dataset range from 0 and 0.4.

# 3.3 Data modelling

## 3.3.1 Environment Set up

Due to a lack of technological infrastructure that could handle such a large dataset, this section details the loading and modeling operations for the movie lens 20 million dataset. As a result of Spark being configured on Google Colaboratory with Graphics Processing Units (GPU) enabled and a link to my Google Drive, where the dataset was kept, infrastructure to handle such datasets could be developed, as illustrated in figures 2 and 3.

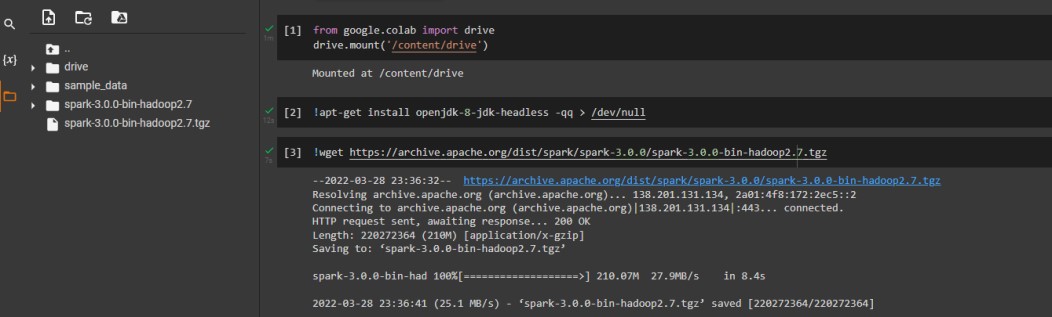


Figure 2: PySpark environment set up

## 3.2 Modelling

In order to model a collaborative filtering system, the Alternating Least Square (ALS) method was used. Important parameters, such as rank, maxIter, and the regularization parameter, were tuned to achieve the best results, with userId, movieId, and rating filling the user, item, and rating columns, respectively. Since the job at hand or the outcome to be predicted is a regression problem, as indicated in figure 3 below, the root means squared score was utilized as the performance indicator to assess the model.

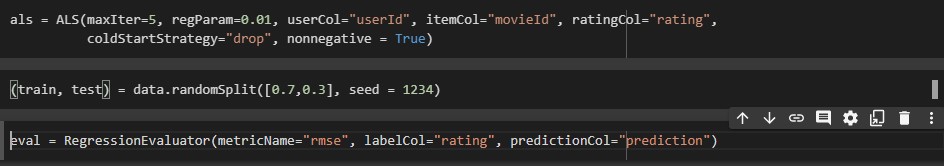


Figure 3: ALS model

It was observed that most predictions failed to predict the rating correctly, checking up the root mean squared error score it was estimated to be 0.8432 with a maximum iteration of 5, regularization parameter of 0.01 and rank of 10, parameter tuning of maximum iteration to 10, regularization to 0.1 and rank to 4 generated a root mean squared error of 0.8352, while tuning of maximum iteration to 15, regularization to 0.2 and rank to 3 increases the root mean squared error to 0.8697 as shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Maximum Iteration | Regularization parameter | rank | RMSE |
| Model | 5 | 0.01 | Default = 10 | 0.8432 |
| Model1 | 10 | 0.1 | 4 | 0.8352 |
| Model2 | 15 | 0.2 | 3 | 0.8697 |

Table 1: Model evaluation

Model 1 was chosen to compute movie rating prediction because it had the lowest RMSE score among the models, as can be seen from the table above. Figure 4 below shows the results of the evaluation of the chosen model on a user.

## 3.3.3 Evaluating Selected Model on a user

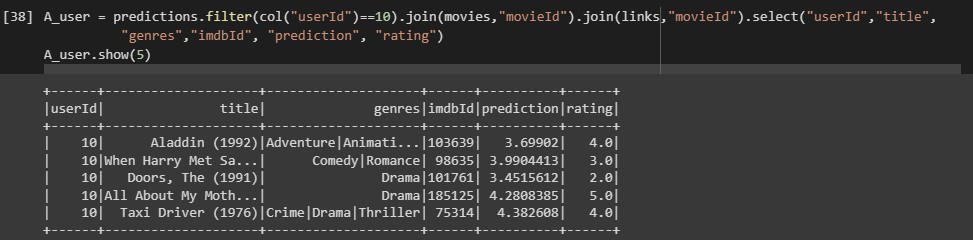


Figure 4: rating prediction

Comparing the projected movie rating to the original movie rating, it was clear that the model accurately anticipated the person with ID=10 movie viewing preferences. Even though the rating projection may not be exact, it may be used to gauge a film's performance before it is released to the general audience.

## 3.3.4 Movie rating simulation

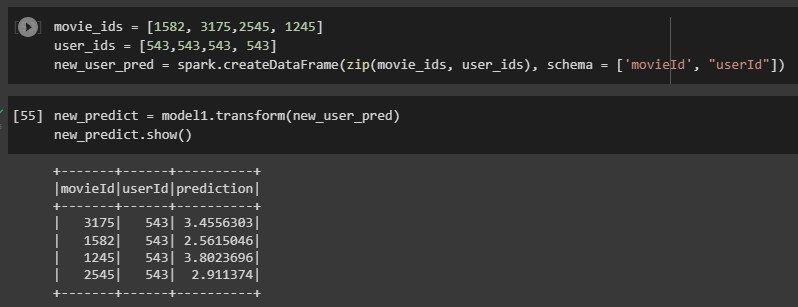


Figure 6: Simulation of recommender system

A recommender system was simulated using the chosen model, with a maximum iteration of 10, regularization parameter of 0.1, and rank of 4. The system was given the user and movie IDs, and it then predicted the rating of the given movie.

# 4.1 Conclusion

This work combined Alternating Least Square machine learning algorithm with a model-based collaborative filtering technique in a Spark environment to construct a recommendation system based on rating using massive data (20+ million) accessible on the Movie Lens platform. As the dataset for the study was cleanly downloaded, the timestamp column in the rating spreadsheet was removed, and as the rating column to be predicted is a continuous variable, the root mean squared error was used to evaluate the model performance. It is clear from Table 1 that model 1 has the lowest RMSE. Model with the lowest regularization parameter (0.01), lowest maximum iteration (5.0), and highest rank (10) has a root mean squared error of 0.8432. Model 1 with a rank = 4, regularization parameter = 0.1 and a maximum iteration = 10 had a root mean squared error of 0.8352, while model 2 with a rank = 3, the highest regularization parameter = 0.2 and maximum iteration = 15 has a root mean squared error of 0.8697, it could be concluded that there is no parameter tuning to ascertain the parameter that improves the performance of the model. This study has shown that the recommendation system is an integral part of the platforms that serve viewers movies for entertainment as it could help increase generated revenue by serving users movies they are yet to watch based on their profile and similarity of watched movie history with other users as shown in figure 4, also it could serve as an integral part of movie production as it helps predict the rating of a movie prior to its release for public consumption as shown in figure 5.

## 4.1.1 Challenges and lessons learned

The main difficulty encountered in this report was set up cross-validation techniques in the modelling aspect. It loaded for about 5 hours before being aborted, forcing manual entry of parameters. If the cross-validation had been successful, time would have been saved and a better model would have been produced. I studied research papers that applied recommendation systems on movies and other products, learned how to handle big data with the Spark analytical tool on Python, and developed the ability to solve big data problems with knowledge of evaluating machine learning algorithms that seek to solve a regression problem. During my research, I came to the conclusion that conventional movie producers could use a recommendation system to increase the percentage of successful films. Success, as was previously stated, is subjective, but considering movie rating should be a normalized form of evaluating movie success, thus the use of rating as a measure of movie success. My interest in the entertainment business has grown as a result of my research, and I want to add to the body of knowledge.

**References**

Ponnam, L.T., Punyasamudram, S.D., Nallagulla, S.N. and Yellamati, S., 2016, February. Movie recommender system using item based collaborative filtering technique. In *2016 International*

*Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)* (pp. 1-5). IEEE.

Deshpande, M. and Karypis, G., 2004. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, *22*(1), pp.143-177.

Wen, Z., 2008. Recommendation system based on collaborative filtering. *CS229 lecture notes*.

Gaenssle, S., Budzinski, O. and Astakhova, D., 2018. Conquering the box office: Factors influencing success of international movies in Russia. *Review of Network Economics*, *17*(4), pp.245-266.

Saraee, M., White, S. and Eccleston, J., 2004. A data mining approach to analysis and prediction of movie ratings. *WIT Transactions On Information And Communication Technologies*, *33*.

Im, D. and Nguyen, M.T., 2011. Predicting box-office success of movies in the US Market. *CS229, Stanford University, Fall*.

Quader, N., Gani, M.O., Chaki, D. and Ali, M.H., 2017, December. A machine learning approach to predict movie box-office success. In *2017 20th International Conference of Computer and Information Technology (ICCIT)* (pp. 1-7). IEEE.