Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

Recommendation of Movies based on Collaborative Filtering using Apache Spark

Rakshitha. P
2nd Semester, M. Tech in CNE,
Dept of ISE, BMSCE, Bangalore

Varsha. R
2nd Semester, MTech in CNE,
Dept of ISE, BMSCE, Bangalore

Dr. Ashok Kumar
Professor Dept of Information Science & Engineering,
BMSCE Bangalore.

***Abstract*:- Nowdays, Recommender Systems (RS) had become more often and trendy as movie provides enhanced entertainment, Movie Recommender (MR) is most important in our social life. Such a Recommender system can suggest a variety of movies to users on the bases of their ratings, interest or the popularities of the movies. In this study we emphasize to execute Recommendation Algorithm using Apache Spark a machine learning tool, in Hadoop File System (HUE) basis to ensure a scalable system to process huge data sets effectively.. The study helps to know the information about maximum ratings along with the count of users who have rated a movie and briefing about all the top most movie prediction for an individual and how often they have rated movie. Eventually the approach is obsessed with the performance of MR mechanism from ALS under various lambda values, iteration further evaluation is performed using RMS (Root Mean Squared) Error of classification forecast which is capable of creating an appropriate rating prediction for movie Recommender.**

Keywords:- Collaborative filtering, Apache Spark, Alternating Least Squares, Recommender System, RMSE, Movielens dataset.

INTRODUCTION

Collaborative filter, compilation of information from vast data collected and to spell out the recommendation. The main reason the recommendation is essential in the present world, is to choose from many options that is available thru the digital media. The attainability is due to availability of internet which allows to access ample resources online. In spite of the fact that the information that is accessible is humongous, while stating that, this lead to a major confusion with many unwanted information being brought down which makes user go unfocused. That is when the recommendation system comes in to effect.

The Recommender System (RS) a system which is effective of foreseeing the subsequent liking of options by the user. It is also a statistical data filtrate system which finds way to forecast the preference or ranking that will be assigned by the user to specific item. Such systems are regularly used by various enterprises. These systems are generally used to generate playlist for music & video services such as YouTube, Netflix, Hotstar or service recommenders on products like Amazon or even recommenders for digital social media such as Twitter & Facebook. We know for sure that there is lot of financial investment on research and development by the entrepreneurs to get the superior techniques to find best possible recommendation to satisfy the customers need & improve on their encounter.

Creating a RS with Spark is very simple a task as its machine learning library does the major & important task for the customer.The user's prediction preferentially, filter collaboratively uses choice by interests similar to other users & try to guess interest of individual taste of movies that are known to the user. To construct recommendation, the Spark MLlib uses Alternate Least Squares (ALS) which is a very popular algorithm for making recommendations.. We should know that to make an ALS occurrence with given parameters one can assign value based on the need. The defaults values are: numBlocks: -1, Iteration: 10, Rank: 10, Alpha: 1.0, lambda: 0.01 and false is the implicitPrefs.

This paper search into a model-based movie recommendation engine, where new users movies are recommended by spark. We can see how ALS interact operate with Matrix Factorisation (MF) for a movie recommendation engine and project uses the movie lens dataset. This paper also gives a very basic knowledge of a standard way of developing RS, Collaborative Filtering.

REVIEW OF LITERATURE

Up until this point, a few analysts presented and introduced research in the region of building Recommendation System in which a current Recommendation calculation can be partitioned into four sorts: content based, Knowledge based, Collaborative Filtering (CF) and Hybrid. In these Recommendation Algorithms, CF is the most well known method, which works by finding past Identical user's interests will share basic interests later on and predicts the ranking of an item dependent on selections & past ranking of the same users.
Here is few related research works associated with collaborative filtering recommender System. Deuk Hee Park et.al. have

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

introduced a characterization and Literature review in their paper and gives understanding about previous work and future extent ofarea.

1. Vikas, Kumar, et al. presents a fresh idea of matrix factorisation and matrix completion for Movielens Dataset in the year 2017 Elsevier by using Collaborative Filtering Approach in order to get the better of the complication of over fitting. This work draws inspiration from an ongoing work on proximal Support vector machines (PSVMs).

2. By using Confidence weighted bias model (CWBM) Technique for online collaborative Filtering (OCF) approach which was proposed by Author Zhou, Xiuze, et al. in the year 2017 for Movielens Dataset to acknowledge continuous updates of proposal results, in order to improve the precision of OCF and to solve the cost of training and improve the stability of OCF in which Trials were directed on Movie- Lens100K and MovieLens1M. This outcome exhibits the lower RMSE values as well as at the same time it is more stable than the other.

3. Li Xie, et al. made an analysis on Collaborative filtering algorithm based on ALS Apache Spark for Movielens Dataset in the year 2017 CIT in order to solve the cold- start problem. By this the root means square of the new algorithm is smaller than that of an algorithm based on ALS in different iterations.

4. Phorasim, Phongsavanh, and Lasheng Yu build up a Movie recommender System utilizing Collaborative Filtering method and K-means clustering algorithm technique for Movielens Dataset in the year 2017 ACCENTS, to enhance data sparsity and scalability.

5. Xuelin Zeng et al. presented a paper on Parallel Latent Group Model (PLGM) technique which was focused on two lattice factorization algorithms were ALS and SGD by using Group Recommendation, Latent Factor approach by selecting Movielens Dataset in the year 2016 to improve the capacity of preparing enormous scope information and to upgrade the quality and versatility.

6. Dianping. Lakshmi et al. utilized item-based collaborative filtering and user item rating matrix technique by using Collaborative filtering approach for Movielens dataset in the year 2016 to improve the scalability, accuracy and data sparsity and error prediction.

7. Zhou, Yunhong, et al. the paper presented in the year 2008 springer which is a direct parallel algorithm for huge scope Collaborative filtering which, on account of the Netflix prize, The model was created using ALS with Weighted Regularization (ALS- WR) to be versatile to large datasets and to improve the better scores of RMSE over Netflix's.

MOVIE RECOMMENDER SYSTEM: A PROPOSAL

The chapter furnishes the proposed system's scheme. By using selected parameters of ALSalgorithm, a better performance of recommender system is been built. The originality of task is based on the preference of framework on ALS method which can influence presentation of structure of a MRS.

a. Block Diagram of RS

While doing this work, we put in ratings of users from the datasets using well know websites like MoviLen, IMDB and TMDb. The availability of dataset in various formats namely databases, CSV file and text file. We have option to download or stream the data live from websites, same is stored either on HDFS or the local file system. The real time data from different origins like geographical system, the stock market and the twitter by using spark system and strong analytics to conduct business, also used in compiling real time streaming of data. Forecasting the grading of users for a specific movie is done by using collaborative filtering (CF) based on the ranking for different movies. Thereafter with another users ranking collaborate for that particular movie. We get the results from machine learning model by training the ALS algorithm using MovieLen data. The Data is stores by using SQL services; spark SQL's data frame and dataset. RDBMS is used to hoard the results of machine learning model; a particular use can retrieve and display recommendation. Local drive is used to store the results of the movie recommendation system.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

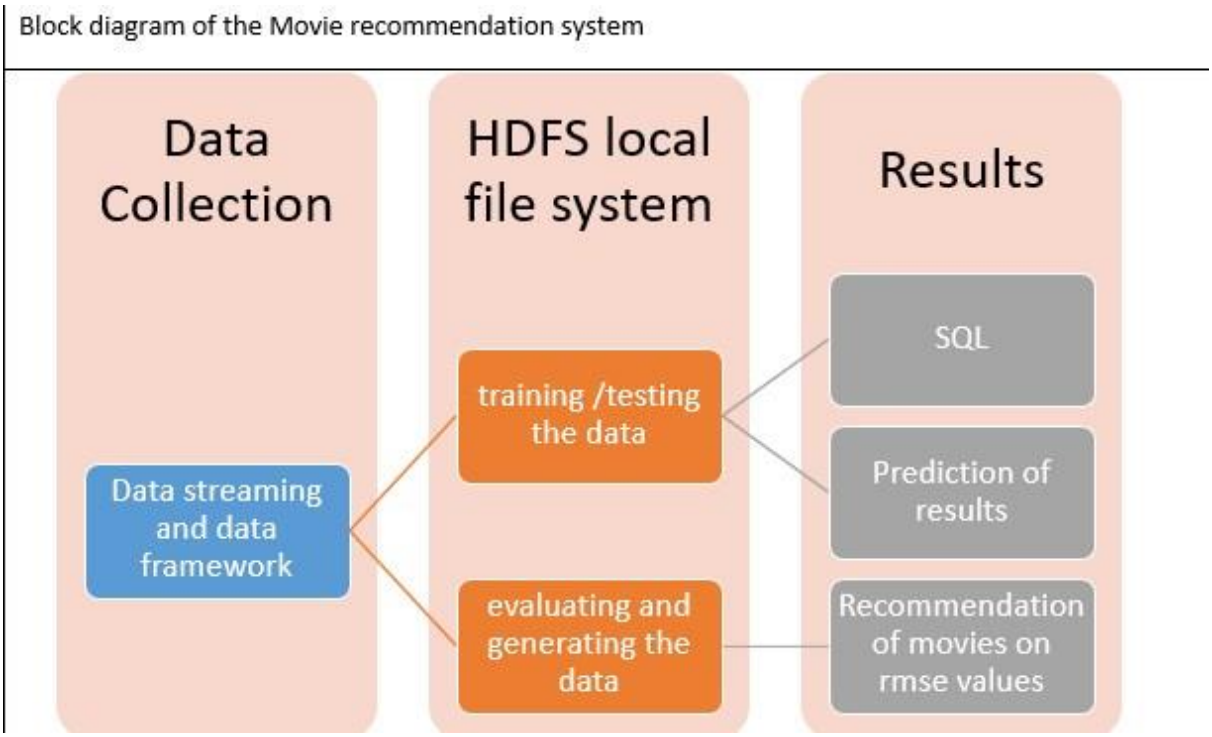


Figure 1:Block diagram of the movie recommendation system.

b. Proposed SystemSteps

This paragraph shows meticulous steps of put in the ALS methods on MovieLens datasets for authenticate choosing of superlative framework while structuring a movie recommendation system.
Recommendation algorithm which is given below shows how the movie lens data set is been taken as the input to the given algorithm and the results is been taken as the output

Recommendation system using ALS and CF

Inputting the movie lens dataset
Outputting the evaluated RS model
Step1=Import the package and loading the dataset
->Storing the dataset of rating.csv
->Displaying the ratings file which has been stored in ratingDf Step 2=The step 1 is been repeated for the movie.csv file.
Step 3=Registration of both the data frame (movieDf,ratingDf) Step 4=Querying and explore the relative dataset
->Total number of users,movies,rating
->Taking the maximum and minimum count of the user ratings
->Most active user schema RRD
->Ratings of most active use (top 10)
Step5=Training and testing the rating data to check the count
Step6=Building an ALS and RS model
Step7=Product matrix
Step8=prediction making
Step 9=To evaluate the RS model

The algorithm will illustrate the ALS algorithm which will have selected parameters which are been selected to have an accurate RS. Here we will have been importing the csv file of movies and rating and the tag files .In which we are storing the csv files and displaying them in the stored top 20 files in the row.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

The next step is were we do registration of the data frames of movies and rating.And the querying and exporting of the data set from movie data set .And from the movie dataset the most active user and their rating with the maximum and the minimum count of the movies are been displayed .
The ALS model is formed with the rmse values which are been raken for diffetent lambda value and the rank of the rmse is 20 and the iteration with different values.And finally the rmse value is been taken with the most accurate RS.From the ALS algorithm.

RESULTS & ANALYSIS

There is an immense growth in popularity of RS. The Apache Spark is used here to demonstrate a well organized aligned execution of a concerted clear algorithm method by using ALS. It is handed-down for proportions depletion cause that help in prevail over restriction of concerted straining for instance data scarce and expandability. The challenge of data insufficiency are become visible in countless circumstances, particularly, complication, one more problem, when a new thing or user has appended to the system, just appended, difficult in finding identical one as there is no adequate particulars, this type of trouble is also known as cold start trouble. While choosing the ALS method as a part of making the suggested MRS, there exist a simple and basic framework thro which can dictate a good classification of customers for the movies given. The particular frame works are Iterations, Rank, and Lambda.

The beneficiation of this article is to review and establish the choosing of frame work that influences the execution of ALS model in structurising a MRS as from the literature study, it is also established that compact research work concentrated on the study of selection of ALS's framework which hamper its execution in constructing a MR mechanism using Apache Spark. The Framework, lambda, & in order to control iterations are used and modify forecast of array factorisation which be contingent on ALS operating procedure this one after the other result the assessment of MRS. Following are used from the lambda and iterations framework: Lambda that describe standardization framework in iterations and ALS in the course of the suggested transcript should run stipulated variety of iterations. ALS method accounts it's ideal explination allying 5 & 25 iterations.

ALS model, that we have chosen with the following parameters of lambda and iteration of different values to check the performance of matrixfactorizationand the better performance of the RS.To illustrate the accurate recommendation system,this will give better performance and results.
The given below tables will represent the better accuracy and the best performance of the movie recommendation system. This will have different lambda values and the iteration values with the differing with movieId and rating of the top 6 movies in the movie RS.

The given fig will represent the maximum and the minimum rating of the top 20 movies and the count of how many have rated the movie.In which we can also see which is the most active user in the given recommendation system.In the figure 1,2 and 3.

The table no 8 and 9 will represent the RMSE values gives the best performance and accuracy of the given RS. When thisis evaluated to the matrix factorization in which we can have the best results.

Case 1
The below table will be showing the iteration value at 10 and for different lambda values the which are been chosen in the given ALS model. Where at the lambda value 0.2 at the given iteration value 10 and the least value is 0.8877, which has a different value of the movieId and rating. RMSE value keeps changing.

lambda	itreations
	10
0.1	0.8918
0.2	**0.8877**
0.3	0.9221
0.4	0.9647
0.5	1.0102
0.6	1.0628
0.7	1.1218
0.8	1.1862

Table 1

iteration=10	
movieId	rating
68945	5.13
96004	5.13
8477	4.98
5490	4.91
13233	4.91
33649	4.85

Table 2

The given table 1 and 2 will show the values of rmse at the iteratin 10 with movieid and rating

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

Case 2

The given below table will represent the iteration value of 15 and for the different values of lambda the values of the iteration keeps on changing.As we have chosen in the given ALS model where the lambda value 0.1 is the least value 0.8718 which has a different value of movieid and rating. The RMSE results are shown below

lambda	itreations	
	10	15
0.1	0.8918	**0.8718**
0.2	**0.8877**	0.8864
0.3	0.9221	0.9204
0.4	0.9647	0.9642
0.5	1.0102	1.0101
0.6	1.0628	1.0627
0.7	1.1218	1.1217
0.8	1.1862	1.1862

Table 3

iteration=15	
movieId	rating
25947	5.08
141718	5.05
5490	5.021
132333	5.021
40491	5
6818	5

Table 4

Representation of the tables 3 and 4 will have the rmse values at iteration 15 with movieid and ratiing

Case 3

The given below tables will be for the iteration 20 which will have a different RMSE value and the given moviid and the rationg also have a different value for the given iteration and the least value is at lambda value 0.2 and the value is 0.886.

lambda	itreations		
	10	15	20
0.1	0.8918	**0.8718**	0.8917
0.2	**0.8877**	0.8864	**0.886**
0.3	0.9221	0.9204	0.9204
0.4	0.9647	0.9642	0.9644
0.5	1.0102	1.0101	1.0102
0.6	1.0628	1.0627	1.0627
0.7	1.1218	1.1217	1.1212
0.8	1.1862	1.1862	1.1854

Table 5

iteration=20	
movieId	rating
26947	5.09
1311718	5.08
5940	5.11
41491	5.11
6818	5.01
14491	5.02

Table 6

The above tables 5 and 6 are the rmse values at iteration 20 and with the differing movieid and rating

Case 4

The representation tables below shows the final RMSE values and the movieid and rating.The given iteration at 25 will be given the RMSE values of lambda is 0.2 at which w ill have the value of the least iteration is 0.886.

lambda	itreations			
	10	15	20	25
0.1	0.8918	**0.8718**	0.8917	0.8917
0.2	**0.8877**	0.8864	**0.886**	**0.886**
0.3	0.9221	0.9204	0.9204	0.9203
0.4	0.9647	0.9642	0.9644	0.9645
0.5	1.0102	1.0101	1.0102	1.0103
0.6	1.0628	1.0627	1.0627	1.0629
0.7	1.1218	1.1217	1.1212	1.1216
0.8	1.1862	1.1862	1.1854	1.1853

Table 7

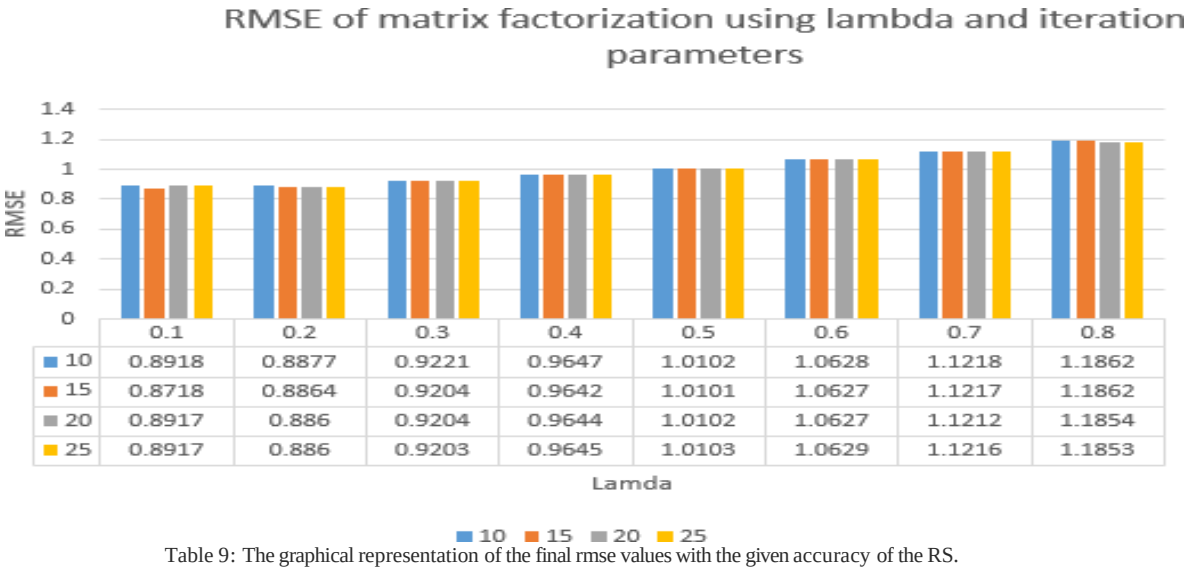
iteration=25	
movieid	rating
5416	4.89
2836	4.89
5328	4.89
39511	4.89
898	4.83
3266	4.81

Table 8

Showing the table 7 and 8 with the final rmse values at the iteration of 25 with different movieid and rating

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

The show RMSE value will be represented as the graphical format



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
10	0.8918	0.8877	0.9221	0.9647	1.0102	1.0628	1.1218	1.1862
15	0.8718	0.8864	0.9204	0.9642	1.0101	1.0627	1.1217	1.1862
20	0.8917	0.886	0.9204	0.9644	1.0102	1.0627	1.1212	1.1854
25	0.8917	0.886	0.9203	0.9645	1.0103	1.0629	1.1216	1.1853

Table 9: The graphical representation of the final rmse values with the given accuracy of the RS.

This will represent the maximum and the minimum and the count of the top 20 rating of the RS in the movie lens database.In which we will have the maximum and the minimum number of movies how have rated and the count of the usersid who has rated those movies.

title	maxr	minr	cntu
Forrest Gump (1994)	5	0.5	329
Shawshank Redemption, The (1994)	5	1	317
Pulp Fiction (1994)	5	0.5	307
Silence of the Lambs, The (1991)	5	0.5	279
Matrix, The (1999)	5	0.5	278
Star Wars: Episode IV - A New Hope (1977)	5	0.5	251
Jurassic Park (1993)	5	0.5	238
Braveheart (1995)	5	0.5	237
Terminator 2: Judgment Day (1991)	5	0.5	224
Schindler's List (1993)	5	0.5	220
Fight Club (1999)	5	0.5	218
Toy Story (1995)	5	0.5	215
Star Wars: Episode V - The Empire Strikes Back (1980)	5	0.5	211
Usual Suspects, The (1995)	5	1	204
American Beauty (1999)	5	0.5	204
Seven (a.k.a. Se7en) (1995)	5	0.5	203
Independence Day (a.k.a. ID4) (1996)	5	0.5	202
Apollo 13 (1995)	5	1	201
Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)	5	0.5	200
Lord of the Rings: The Fellowship of the Ring, The (2001)	5	0.5	198

only showing top 20 rows

FIGURE 2 : Represents top 20 maximum and minimum rating with the count of users

The five most active users are shown below by the movie lens dataset which have been used for the given analysis.The analysis shows the 414 user id is the most active and the count of the movie which are been rated by them.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

```
scala> val mostActiveUsersSchemaRDD = spark.sql("SELECT ratings.userId, count(*) as ct from ratings "+ "group by ratings.userId order by ct desc limit 10")
mostActiveUsersSchemaRDD: org.apache.spark.sql.DataFrame = [userId: string, ct: bigint]

scala> mostActiveUsersSchemaRDD.show(false)
+------+----+
|userId|ct  |
+------+----+
|414   |2698|
|599   |2478|
|474   |2108|
|448   |1864|
|274   |1346|
|610   |1302|
|68    |1260|
|380   |1218|
|606   |1115|
|288   |1055|
+------+----+
```

FIGURE 3: The most active users with user id and the count of movie rating

This will show the user, how most active on the movie lens data set is 414 which will display the top 20 movies and the rating which have been given by the user id 414 and the movies name which are been rated by him.
FIGURE 4: Most active user id 414 with top 20 movies which the user has rated

5. CONCLUSION

```
scala> results2.show(false)
+------+-------+------+----------------------------------------------------------------------
|userId|movieId|rating|title
+------+-------+------+----------------------------------------------------------------------
|414   |11     |5     |American President, The (1995)
|414   |32     |5     |Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
|414   |34     |5     |Babe (1995)
|414   |50     |5     |Usual Suspects, The (1995)
|414   |94     |5     |Beautiful Girls (1996)
|414   |110    |5     |Braveheart (1995)
|414   |111    |5     |Taxi Driver (1976)
|414   |151    |5     |Rob Roy (1995)
|414   |223    |5     |Clerks (1994)
|414   |260    |5     |Star Wars: Episode IV - A New Hope (1977)
|414   |266    |5     |Legends of the Fall (1994)
|414   |290    |5     |Once Were Warriors (1994)
|414   |293    |5     |Léon: The Professional (a.k.a. The Professional) (Léon) (1994)
|414   |296    |5     |Pulp Fiction (1994)
|414   |318    |5     |Shawshank Redemption, The (1994)
|414   |322    |5     |Swimming with Sharks (1995)
|414   |353    |5     |Crow, The (1994)
|414   |356    |5     |Forrest Gump (1994)
|414   |417    |5     |Barcelona (1994)
|414   |431    |5     |Carlito's Way (1993)
+------+-------+------+----------------------------------------------------------------------
only showing top 20 rows
```

Movie recommender device play a widespread position in identifying fixed movies for customers, primarily on customer's interest. Even though number of flow RS are made accessible for customers. In this article provided a film influencer gadget based totally on collaborative filtering the usage of Apache Spark. At the outcome, the choice of framework of ALS algorithm will have an effect on the execution of constructing a film influencer platform. Methodology assessment is made, the use of number of measures together with implementation schedule, RMSE of classification forecast. Movie Recommender System Based on Collaborative Filtering model turned into training. First couples of premier manifestation are opted based on fine framework choice from inventive consequences; this could result in constructing quality forecast score for a MR mechanism. Through such cases, the least fee of RMSE is observed as excellent occurrence for forecast while constructing MR device. Finally we here by conclude that the RMSE value of the lowest will have an accurate RS from the movie lens dataset.

REFERENCES:

1. https://www.researchgate.net/profile/Mohammed_Fadhel_Aljunid/publication/335378751_A_SURVEY_ON_RECOMMENDATION_SYSTEMS_FOR_SOCIAL_MEDIA_USING_BIG_DATA_ANALYTICS/links/5d6113ad458515d61020cdb8/A- SURVEY-ON-RECOMMENDATION-SYSTEMS-FOR-SOCIAL-MEDIA-USING- BIG-DATA-ANALYTICS.pdf
2. https://link.springer.com/chapter/10.1007/978-981-13-1274-8_22

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

3. https://mapr.com/ebooks/spark/08-recommendation-engine-spark.html
4. https://spark.apache.org/docs/2.2.0/ml-collaborative-filtering.html
5. https://medium.com/@navdeepsingh_2336/scala-machine-learning-projects- recommendation-systems-d41d9eebbb06
6. https://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/recommendation/ALS.html
7. https://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/Estimator.html
8. https://towardsdatascience.com/prototyping-a-recommender-system-step-by-step- part-2-alternating-least-square-als-matrix-4a76c58714a1
9. Movie Recommender System Based on Collaborative Filtering Using Apache Spark Mohammed Fadhel Aljunid and D. H.Manjaiah
10. Evaluating and Enhancing Efficiency of Recommendation System using Big Data Analatic. Archit Verma, Dharmendra Kumar.
11. Performance Analysis of Various Recommendation Algorithms Using Apache Hadoop and Mahout Dr. Senthil Kumar Thangavel, Neetha Susan Thampi, Johnpaul CI
12. Kumar, Vikas, et al. "Collaborative filtering using multiple binary maximum margin matrix factorizations." Information Sciences 380 (2017): 1-11
13. Zhou, Xiuze, et al. "Confidence-weighted bias model for online collaborative filtering." Applied Soft Computing (2017).
14. Xie, Li, Wenbo Zhou, and Yaosen Li. "Application of Improved Recommendation System Based on Spark Platform in Big Data Analysis." Cybernetics and Information Technologies 16.6 (2016): 245-255.
15. Phorasim, Phongsavanh, and Lasheng Yu. "Movies recommendation system using collaborative filtering and k-means." International Journal of Advanced Computer Research 7.29 (2017): 52.
16. Zeng, Xuelin, et al. "Parallelization of Latent Group Model for Group Recommendation Algorithm." Data Science in Cyberspace (DSC), IEEE International Conference on. IEEE, 2016.
17. Ponnam, Lakshmi Tharun, et al. "Movie recommender system using item based collaborative filtering technique." Emerging Trends in Engineering, Technology and Science (ICETETS), International Conference on. IEEE, 2016.
18. Zhou, Yunhong, et al. "Large-scale parallel collaborative filtering for the netflix prize." Lecture Notes in Computer Science 5034 (2008): 337-348.