Data Analytics

Assignment -8

Name : R Siva Girish SRN : PES1201700159

Dataset : MovieLense inbuilt R Dataset

Recommendation Systems

A Recommendation system is a data filtering tool that makes use of the data collected from other users to recommend the most relevant items to a particular user. Small scale shops may not have the need for a recommendation system but big online retailers need one due to the fact that their catalogue of items is very huge and cannot be displayed all at once.So a good recommendation system is the way to solve this problem.

User Based Collaborative Filtering

This method works by finding the similarities between all the individual users and the actual user to recommend.Based on the data collected it assigns the desired rating to be the weighted average of the ratings given by each of the similar users.

Objective

Build a User Based collaborative filtering Recommendation system on the movie Lens Data.

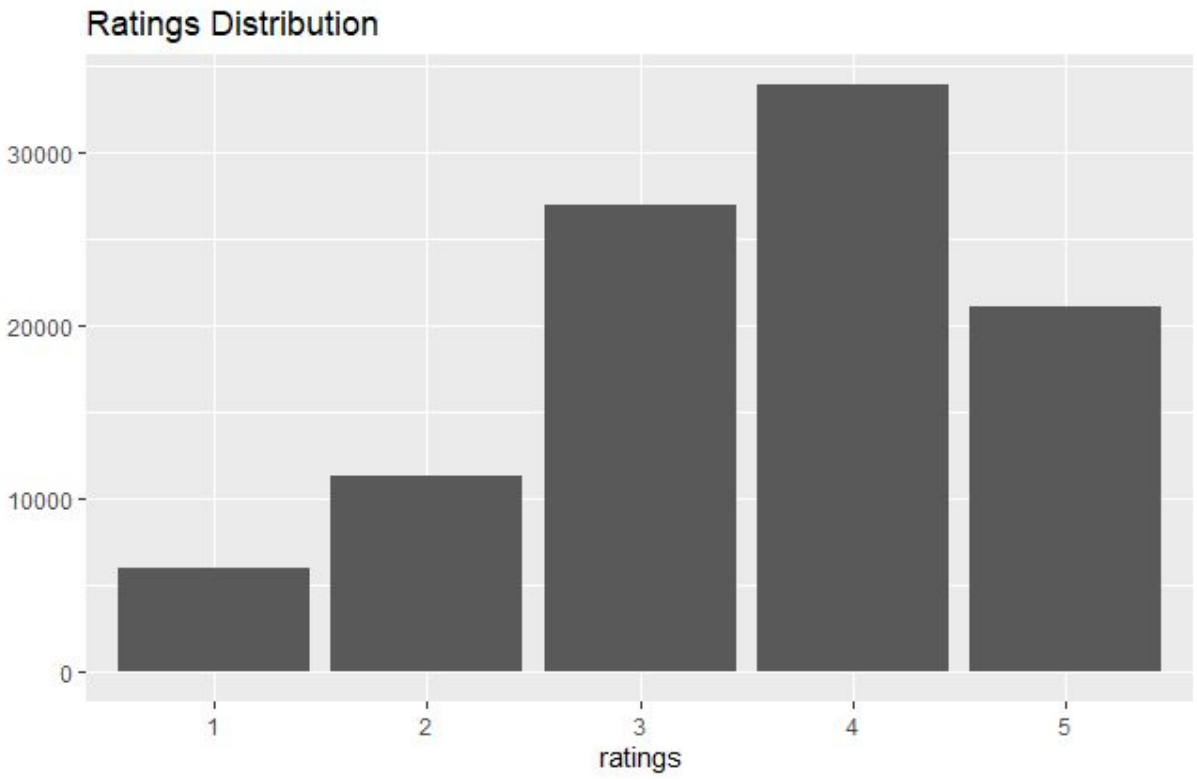
Exploratory Data Analysis(EDA)

- ❖ Find all the unique ratings and their number of occurrences.

```
ratings <- as.vector(MovieLense@data)
table_r <- table(ratings)
table_r
```

```
ratings
     0      1      2      3      4      5
1469760   6059  11307  27002  33947  21077
>
```

```
qplot(ratings) + ggtitle("Ratings Distribution")
```



From the above graph it can be observed that most of the ratings are above 2 and the most common rating given by user in general is 4.The count of users giving rating 4 is 33947.

Data Preprocessing

Preprocessing involves selecting the relevant data as well as normalizing it.

- ❖ Selecting Relevant Data

  We perform this operation by selecting only the users who have rated at least 50 movies (i.e to determine whether they are serious critics or not) and the movies that have been watched at least 100 times (i.e to make sure that the movie is not any rare movie which is liked only by a very small group of people).

```
ratings_movies <- MovieLense[rowCounts(MovieLense) > 50,colCounts(MovieLense) > 100]
```

Splitting into Training And Test Set

Randomly split the data into training and testing dataset which can be done by the 80 20 split.We split the dataset into training and testing using the 80 -20 split.

```
trainer <- sample(x = c(TRUE, FALSE), size = nrow(ratings_movies),replace = TRUE, prob = c(0.8, 0.2))
head(trainer)
recc_train <- ratings_movies[trainer, ]
recc_test <- ratings_movies[!trainer, ]
```

Once the dataset has been split successfully all we must do now is to train our recommendation system and test on the testing set.

Recommendation Model

Now we can build the user based collaborative filtering model.

```
recc_model <- Recommender(data = recc_train, method = "UBCF")
recc_model
```

Let us Recommend top 6 items to the users in the testing set

```
n_recommended <- 6
recc_predicted <- predict(object = recc_model,newdata = recc_test, n = n_recommended)
recc_matrix <- sapply(recc_predicted@items, function(x){
 colnames(ratings_movies)[x]
})
```

```
> recc_matrix[, 1:4]
     14                              25                              58
[1,] "Contact (1997)"                "Good Will Hunting (1997)"      "L.A. Confidential (1997)"
[2,] "Wrong Trousers, The (1993)"    "L.A. Confidential (1997)"      "Leaving Las Vegas (1995)"
[3,] "Jackie Brown (1997)"           "Titanic (1997)"                "Jackie Brown (1997)"
[4,] "Casablanca (1942)"             "Apt Pupil (1998)"              "Apt Pupil (1998)"
[5,] "Forrest Gump (1994)"           "Lawrence of Arabia (1962)"     "Contact (1997)"
[6,] "Wag the Dog (1997)"            "Jackie Brown (1997)"           "Donnie Brasco (1997)"
     75
[1,] "Amadeus (1984)"
[2,] "Monty Python and the Holy Grail (1974)"
[3,] "Schindler's List (1993)"
[4,] "L.A. Confidential (1997)"
[5,] "Good Will Hunting (1997)"
[6,] "One Flew Over the Cuckoo's Nest (1975)"
>
```

The above picture gives us the data of the top 6 recommendations to be given to the first four users in out testing set based on their similarity with the users on the training set.